Machine Learning Engineer Nanodegree

Capstone Project

Rachael Mahon

June 2018

I. Definition

Project Overview

The intention of this project is to explore UK criminal custodial sentencing data in order to evaluate whether predictions can be made about the duration of custodial sentence based features of the crime and features of the defendant.

To do this, I will evaluate the performance and predictive power of a model that has been trained and tested on data made available by the British Ministry for Justice in November 2011. The Ministry released 1.2 million records of criminal sentencing data for the majority of courts in England and Wales. The dataset is anonymised but does contain the age ranges, sex and ethnicities of the subjects. It also contains the sentencing court, the type of offence and the police force who dealt with the matter. The dataset also includes, which is why I am chiefly interested, the ultimate sentence handed down to the defendant.

To achieve this stated aim, I will construct a working model which has the capability of predicting the sentence. I will separate the data into features and a target variable. I will need to convert the features and target variable into continuous values prior to training the model. I have decided to use a multi dimensional decision tree regressor to do this. I will use a single decision tree regressor and linear regressor as my benchmark models.

The intention of this project is not to create a method of suggested durations of sentence. Some companies have created software for the purposes of recommending sentences to judges based on algorithms which, in my opinion, violate the right to due process as defendants and their advocates are unable to scrutinise or challenge the algorithm due to it's protection by intellectual property law. Furthermore, they are based on historically sentence and are so destined to repeat the biases already ingrained in the criminal justice system.

Cathy O'Neil, author of Weapons of Math Destruction, has roundly pointed out how problematic the use of these completely unknown and sealed algorithms are, for example in the area of predictive policing software. These predictive models should not be treated as neutral as they are clearly influenced by the systemically problematic underpinning data and the goals and ideology of those who create and commission them.

Problem Statement

When accused of a crime, you do not have enough data to make decisions about whether you should plead guilty or go to trial, if you need to make arrangements about your employment, property or child care or whether you should appeal a sentence you feel is too harsh. There are too many unknowns and it is very difficult to get an accurate answer on these things.

If the problem is that the criminal justice system is stressful and alienating for many people being processed by it, the solution would be to use previous data to provide some insight for managing expectations or decision-making on whether or not to go to trial or to appeal a sentence. The output when this project is completed should be the capacity for someone to plug in some their personal details, the offence, the court etc, and receive a reasonably accurate sentencing figure based on a regression analysis.

Metrics

R2

The metric I have chosen to use for each stage of this linear regression analysis is R2. It returns a value between 0 and 1 where 0 indicates that the model is not explaining any of the variability in the data and 1 indicates that the model explains exactly all the variability in the data ie. the higher the R2 figure, the better. It is calculated as 1 - the residual sum of square / total sum of squares.

Using R2 as our metric has an interesting quality in that R2 may be negative when applied to unseen data. It is very possible that this may be the case in our data as there are a few outliers and if they happen to fall into the testing data on shuffle split, R2 may be negative.

II. Analysis

Data Exploration

Overview

The data comprises of sentencing data for the majority of courts in Wales and England. The dataset is anonymised but does contain the age ranges, sex and ethnicities of the subjects. It also contains the sentencing court, the type of offence and the police force who dealt with the matter. The dataset also includes, which is why I am chiefly interested, the ultimate sentence handed down to the defendant.

Age - categorical - Age of the defendant split into three categories: 18-24, 25-34 and 35+
Court - categorical - The Court that heard the cases
Offence Type - categorical - The category of offence split into 15 categories
Sentencing Outcome - categorical - There are a number of outcomes but we are only interested in those that resulted in the immediate custody of the defendant. All other categories, such as fines or unknowns are removed at the pre-processing stage.
Amount - categorical - The duration of the sentence split into 12 categories
Force - categorical - The Police force responsible for the cases
Sex - categorical - the sex of the defendant - male, female or not stated
Ethnicity - categorical - The Ethnicity of the defendant - white, black, Asian, unknown or other
Offence to completion - categorical - the time taken from the offence date to the sentencing date

Because the dataset contains only categorical variables, some preprocessing will be done to turn them into continuous variables.

Supplementary Data

I have also made use of the Crime Severity Score data tool released by the Office of National Statistics in June 2017. This data is described by the Office of National Statistics as a:

"list of weights as a reflection of the legislation set by Parliament on behalf of the public and the courts in passing sentences in line with legislation and sentencing guidelines. It is not intended to be a pure ranking of severity of offences; it provides the basis for deriving a Severity Score rather than comparing weights for individual offences."

For each offence type, I have created an average "weight" by getting the average of all the offences listed in that category from this list. It is a crude measurement of severity of the crime given the Office of National Statistics quote above and the means by which I have gotten it but we can still use it to explore the correlation between these weights and sentence duration. The results of this calculation are visible in file 'offence_severity.txt'.

The data is available here:
https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/datasets/crimeseverityscoredatatool

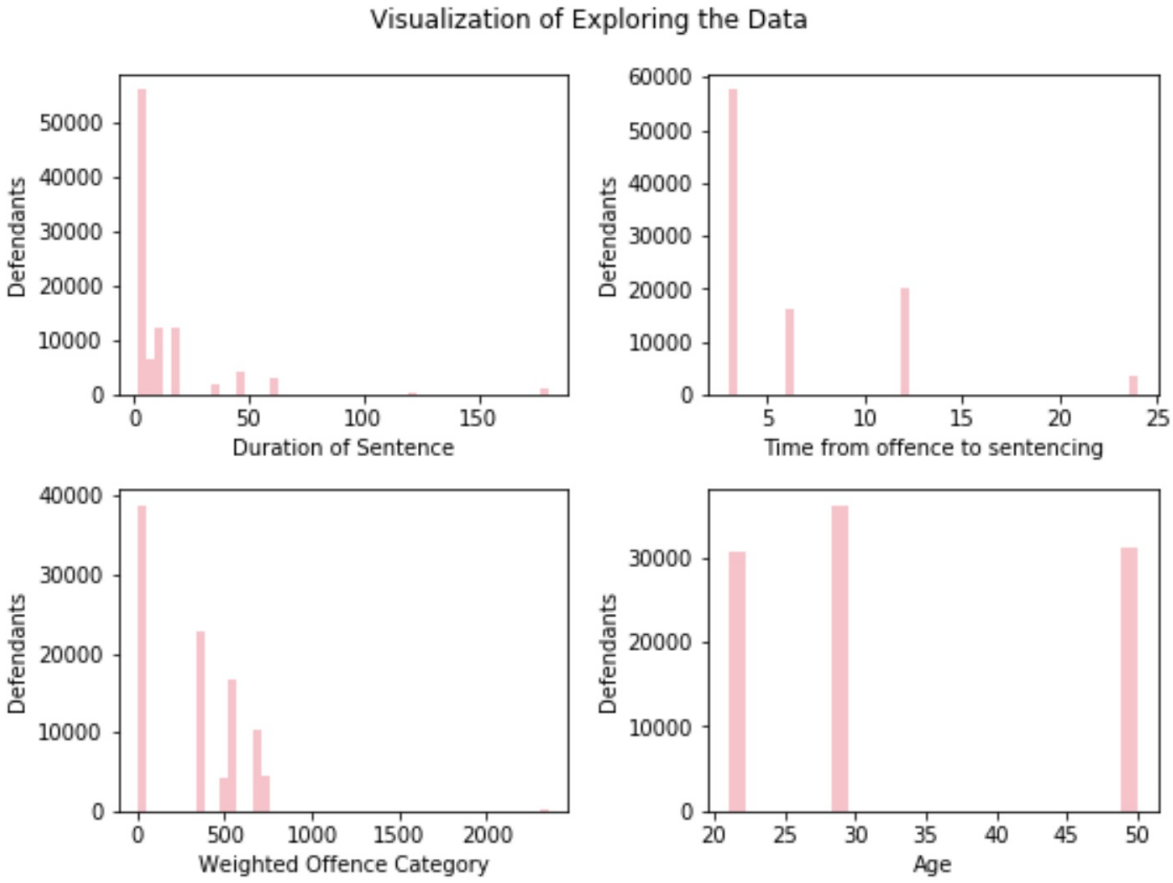
Outliers

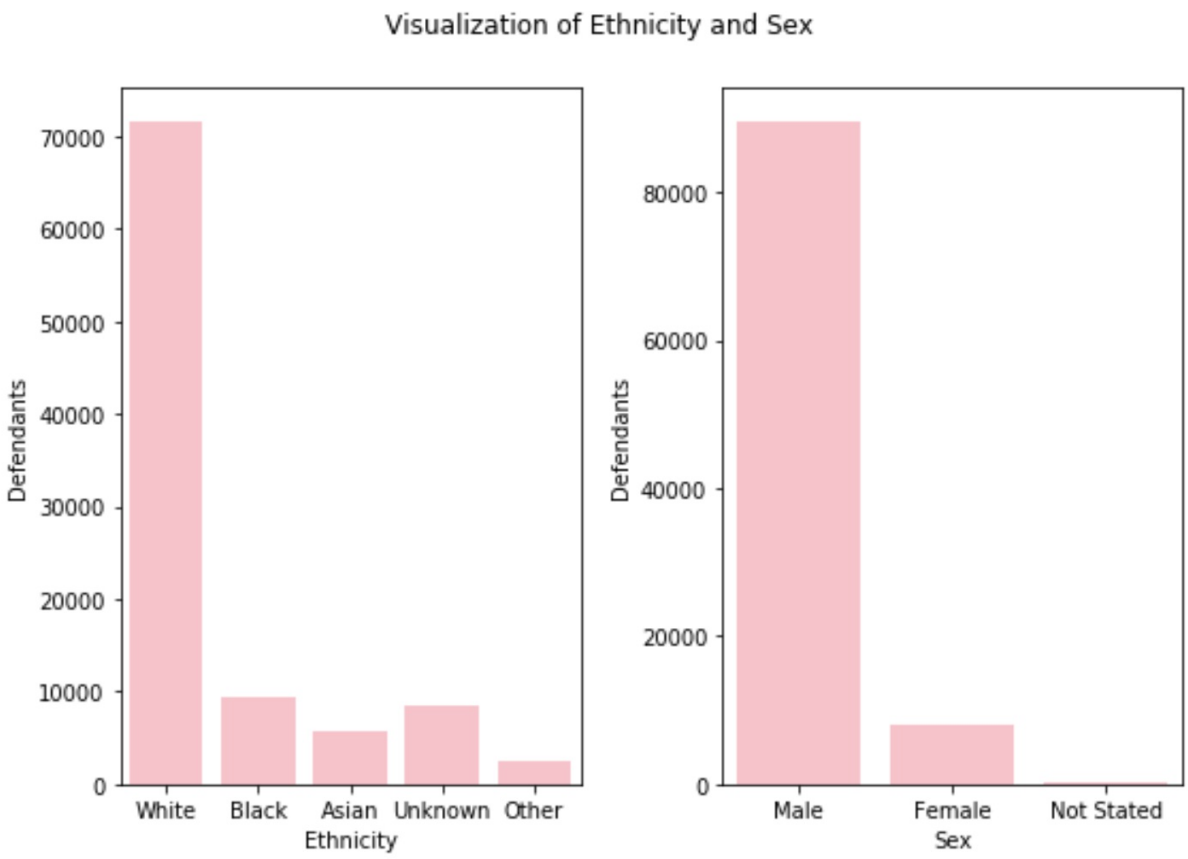
Upon exploring the data, I chose not to remove any outliers for two reasons. Firstly, due to the nature of the data, it is unlikely that there are an incorrect entries in the data. I trust that the data presented by the Ministry of Justice is accurate. Furthermore, the data is mostly categorical and has been altered into pseudo-continuous variables using dictionaries. Because of this, there are big steps in the data and it is very prone to bunching around small sentences with a few very big sentences. Because of this, I do not believe that the removal of a particularly extreme sentence will be helpful. However, it is considerably effecting the results for predictions. This will be discussed further at the evaluation section.

The Data at a Glance

- There are 97814 records
- Minimum sentence: 2 months
- Maximum sentence: 180 months
- Mean sentence: 12.9878340524 months
- Median sentence 3.0 months
- Standard deviation: 24.8257878998
- The full dataset is 97814 rows x 9 columns

Exploratory Visualization





Visual Explorations of the Data

- Age
  The age categories in which people were convicted of offences were quite evenly spread amoung the three categories but was slightly higher in the 25-34 category. It seems that convictions drop off significantly after the age of 35 as there are fewer people in the entire bracket of 35+ than there are in 25-34 and about the same as is in 18-24.

- Offence Category
  There were far more people being convicted of lower level offences and the more serious the offence, the fewer convictions there were. This may mean that there were fewer of these types of offences or that they are harder to prove beyond reasonable doubt and defendants are less likely to plead guilty to them.

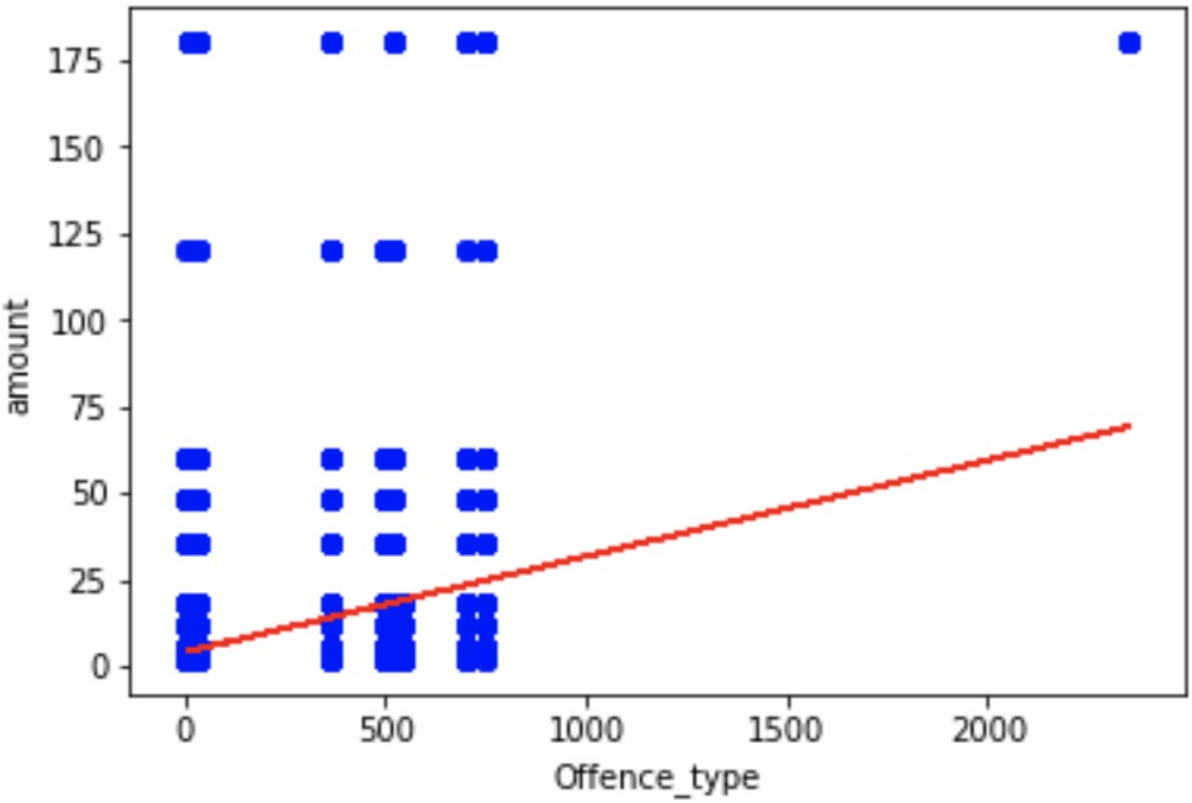
- Offence to Completion
  Most defendants were processed by the criminal justice system relatively quickly. Most cases moved from offence to completion in less than 6 months.

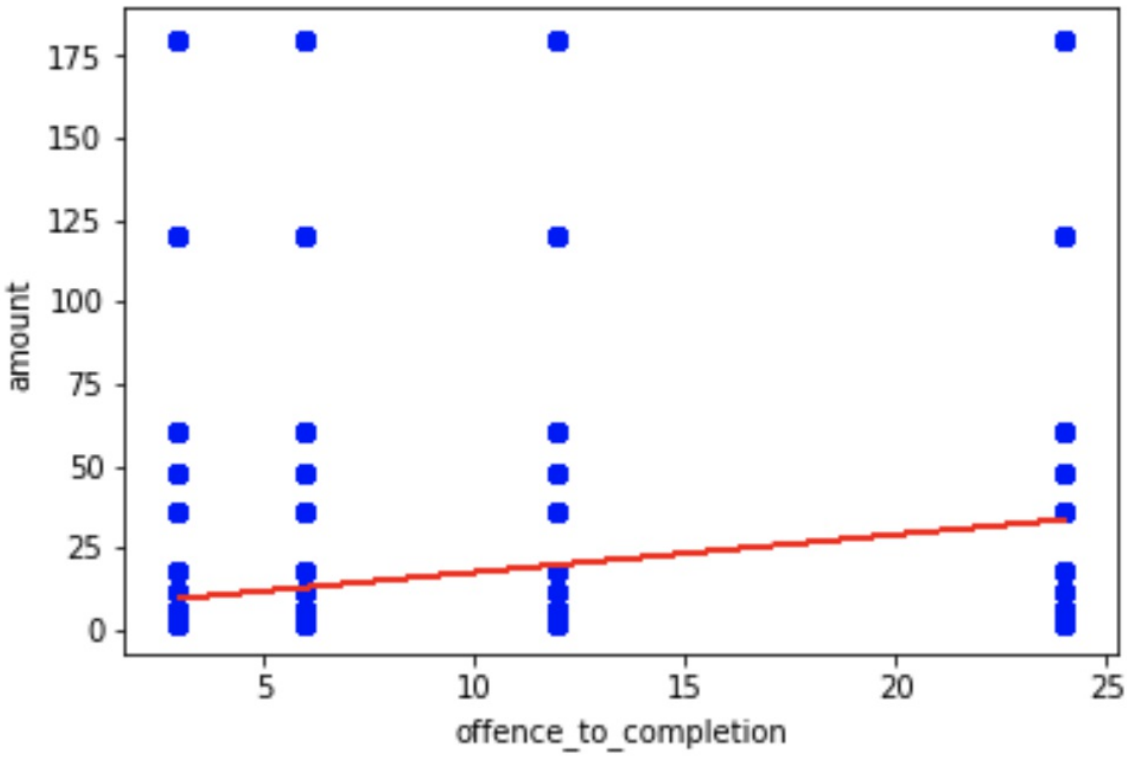
- Duration of Sentence
  A large amount more defendants were sent to prison for very short sentences of "up to and including 3 months." This type of sentence can be very disruptive to a persons life. This will be discussed further in the report.

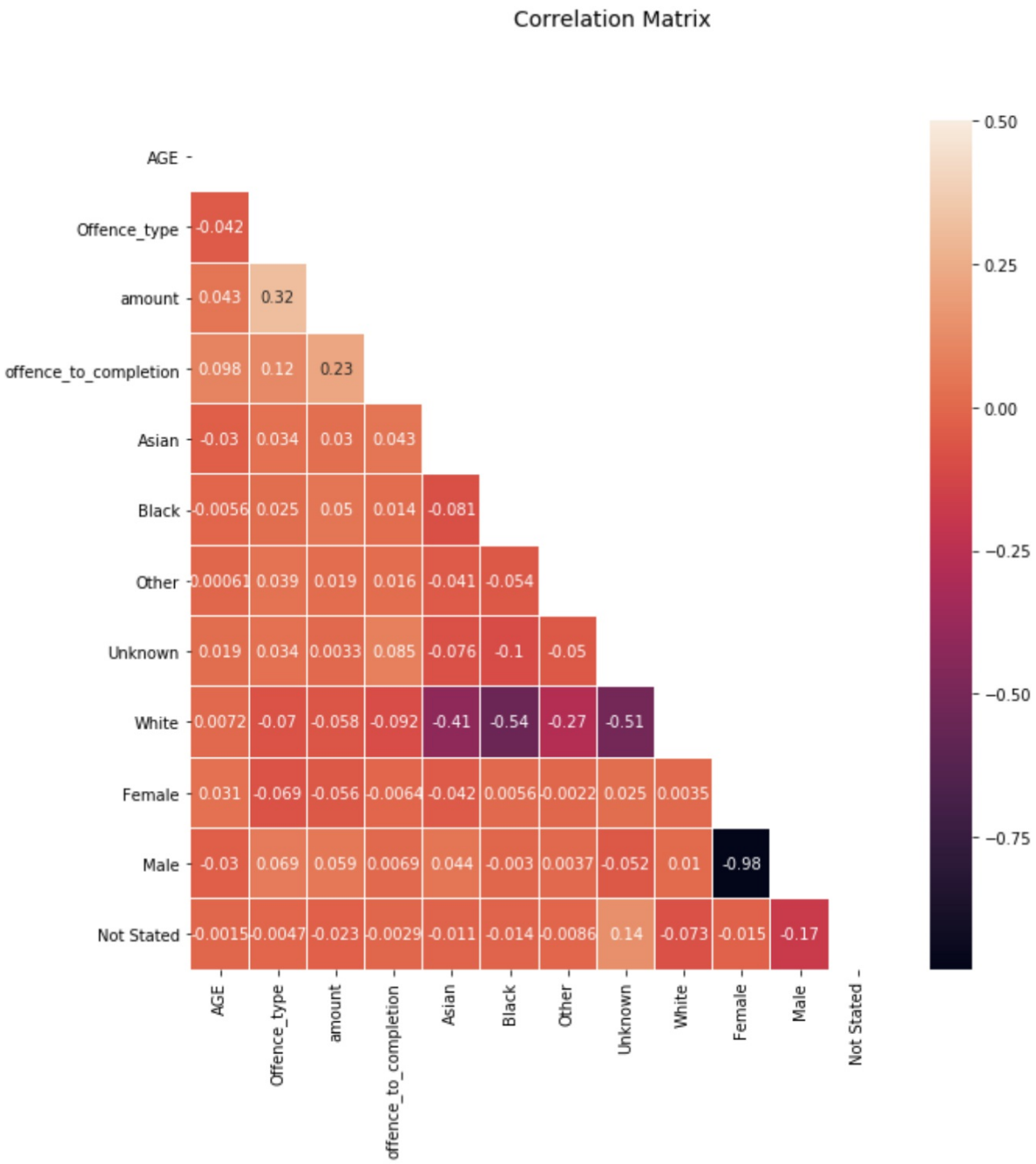
- Ethnicity
  A much higher proportion of ethnically white people were convicted of crimes and sentenced to custodial sentences than any other ethnicity.

- Sex
  A much higher proportion of men received custodial sentences than women.

Correlation Explorations




175
150
125
100
75
50
25
0
amount
5
10
15
20
25
offence_to_completion



As a sanity check, I am pleased that being male correlates so strongly with not being female and being white correlates strongly with not being black.

As expected the duration of the sentence appears to correlate strongly with the weighted severity of the offence.

Surprisingly (to me at least), but understandably, the duration of the sentence appears to correlate most with the length of time between the offence and sentencing. This may be because not guilty pleas (and so trials) receive harsher sentences or because defendants are more likely to enter a not guilty plea on an offence with a long sentence or because the more severe the crime, the more complex the case and so the longer it will take to reach ultimate

sentencing.

Algorithms and Techniques

To implement this, I followed the below steps:

1. Data preprocessing and converting of continuous variables
2. Exploring the data and considering what might be significant features
3. Validating some theories using correlation matrix
4. Removing any unneeded dimensionality
5. Splitting the data with cross validation
6. Defining a performance metric
7. Creating bench mark models
8. Using grid search to find the optimal parameter
9. Validating that this is the optimal parameter using learning curves and complexity models
10. Creating the decision tree regressor and fit it with our training data - this was by far the easiest portion of the task. It worked totally out of the box. It seemed so easy that I assumed something must be wrong for a long time and was trying to debug it before I realised that it actually did just work out of the box like that.
11. I then evaluated the performance
12. And finally fed it some data to return predictions on which seemed reasonable.

There were no particularly difficult parts to code. To be honest, a lot of the main difficulties I encountered related to my versions of python, deprecated libraries and incompatible imports that required a lot of debugging.

Linear regression

I chose to use a simple linear regressor as an initial bench mark model to conveniently visualise and conceptualise the data and problem. Linear regression makes for a nice benchmark model for this type of task that is a dependent variables relationship to an explanatory variable. It can only model one relationship so it is key to select the correct variable to demonstrate the relationship. This is selected by finding the most highly correlated variable using techniques like a correlation matrix.

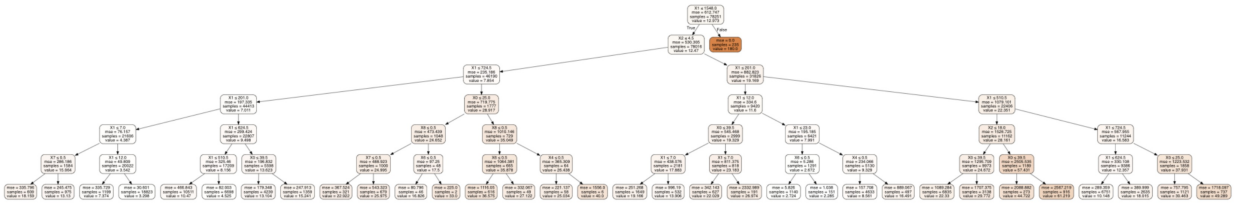
Decision Trees

Decision tree regressors breaks the data down into smaller and smaller subsets based on learning a set of decision rules. It is slightly difficult to see fully but below is an image of the full decision tree ultimately used by the model. There are non-leaf nodes or decision nodes which are mean squared error to make a decision on the split ie a test on attributes. Each branch from the root node is a representation of the outcomes of tests. The nodes at the end of each

branch are the leaf nodes and represent the class that a piece of test data will be characterised as if it arrives at that node. The max depth of these branches is set as the length required to reach the optimal result beyond which the model may cease to be able to generalise on unseen data but is deep enough that it adequately represents the training data.

It can model more complicated, multi dimensional relationships which linear regression cannot. A single decision tree regressor can also outperform a linear regressor on data with influential outliers such as this data.

I employed the use of both a single decision tree regressor (with the most correlated feature) and multiple decision tree regressor (retaining multiple features for the training data).



Decision Tree Regression with one parameter but with different depths

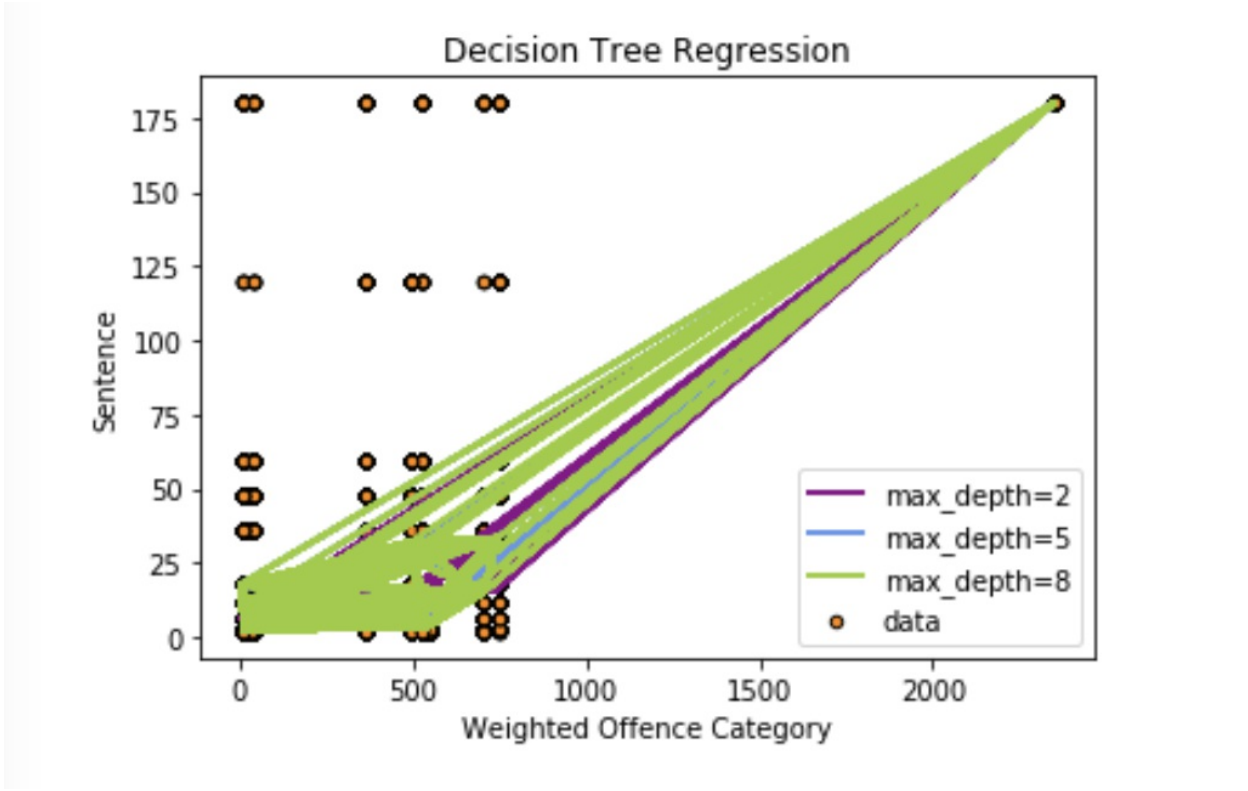
I have also chosen to explore the possibility of a single decision tree regressor around the most strongly correlated feature. I wanted to explore the performance of a decision tree regressor with one feature versus a decision tree regressor with more dimensionality.

Decision Tree Regression with multiple parameters

I chose to use a decision tree regressor because I felt that this algorithm is intuitive when using data with continuous variables and binary variables. It is also convenient for conceptualising the problem.

Benchmark

As explained above, I undertook two bench mark models. Firstly a linear regression which performed very poorly (R2 of 0.10) and three decision tree regressors with max depths of 2, 5 and 8, all of which performed very slightly better than my simple linear regressor ie R2 of 0.18, 0.25 and 0.26 respectively.



III. Methodology

Aside from the problems with the data discussed at length elsewhere, creating this model was very straightforward and I followed the below process. Decision tree regressors work pretty much out of the box so below the enumerated steps are some comments about data processing and optimising the hyperparameter. These are to aid any one wishing to replicate these results.

Data Preprocessing

For the purposes of this project, we are only interested in custodial sentences ie. when a defendant was actually deprived of their liberty. This data additionally contains fine data which is also a very interesting area of study but unfortunately goes beyond the scope of this project. Removing records where the defendant was sentenced to only a fine or where the sentence duration is missing for some other reason removes a very large number of records from the dataset. While it is not ideal to reduce the dataset by this amount, the original dataset was so large as to be unwieldy on my personal machine and I would have had difficulty processing it in depth. The original data contained records and the data with fines and empty sentences removed has 97814 records.

Dictionaries to Convert Categorical Variables to Continuous

To convert the categorical data into continuous variables, I created a number of dictionaries

which are explained below. They are crude measures but this was required to work with the data as more nuanced data is not available.

1. Age Groups

This only has three distinct categories so I am using the average age of all the ages in each category. For 35+ I am assuming the top record is 65. This is an arbitrary distinction.

2. Offence to Completion in Months

I wanted to retain this variable as I believe it is likely highly correlated with the severity or complexity of the crime, whether or not the defendant pleaded guilty (a longer time to completion may indicate a guilty plea and a full trial). Because of this I converted the categorical to the minimum amount in months indicated by the string.

3. Offence Severity

The dataset used to get the weighting of each crime is very interesting and available here:

https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/datasets/crimeseverityscoredatatool

To create the dictionary used to turn the type of crime from a distinct category into a weighted indicator of severity, I used the Crime Severity Score data tool released by the Office of National Statistics as discussed above.

For each offence type, I have created an average "weight" by getting the average of all the offences in that category in the supplementary dataset.

4. Amount

Similarly, for amount, I used a dictionary created from mapping the categories to the minimum number of months to fit into the category.

Encoding Other Categories to Binary Variables

I assume that the dimensionality gained by retaining sex and ethnicity as features will be useful so I encoded these features to binaries variables. I think sex is very unlikely to a useful as the number of women represented in the data is almost vanishingly small. Similarly, ethnicities that are non-white are not represented very much in the data but these are significantly larger than the proportion of women. I retained these so that I could explore if my assumptions are correct.

Implementation

Below is an enumerated list of how I completed the actual implementation after all the data was processed, bench mark models completed and metrics decided upon:

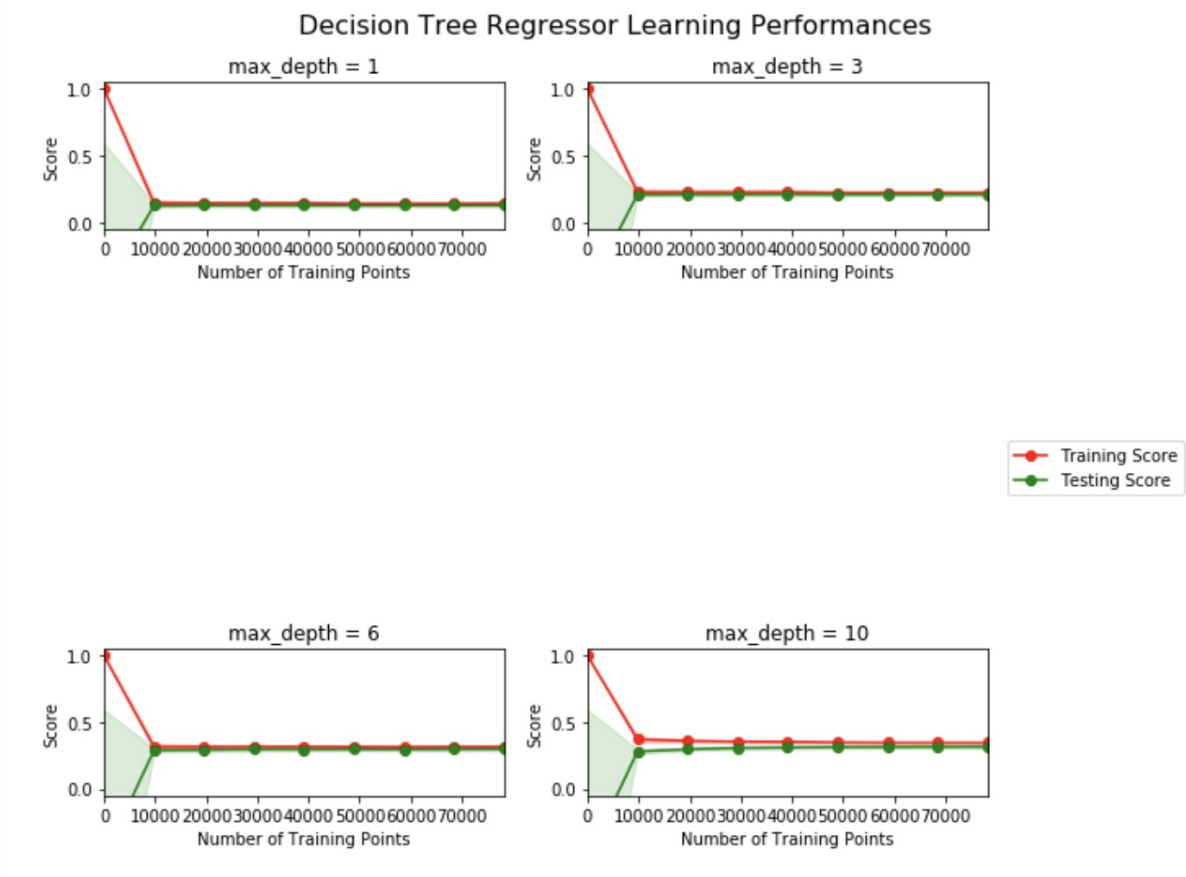
1. Using grid search to find the optimal parameter

To do this, you create a new Decision Tree regressor from sklearn, a new scoring function using make_scorer from sklearn metrics and a new grid from the GridSearchCV library. The regressor, a list of finite parameters to check, the scoring function (which is our overall performance metric in this case) and some cross validation sets are passed to the grid search as parameters. This then trains the model for each parameter in the list on the cross validated training sets and determines which one is the best based on the performance metric that it was supplied with.

2. Validating that this is the optimal parameter using learning curves and complexity models

I reused some code provided during the Boston Housing project earlier on in the Nanodegree. This was for the purpose of visualising the selection of the optimal hyperparameter. This was very helpful.

We can see from the below learning curve diagrams that the optimal max depth for our model is 6. The training score and testing score here are perfectly aligned. It seems very close for max depth one and three also but there appears to be a divergence at the learning curve of max depth of 10. This finding is validate further below in our R2 scores for different max depths.



The model complexity graph below also validates these findings. We can see the validation and training scores being very tight with each other and slowly beginning to diverge after a max depth of 8.

After a max depth of 8 the model begins to suffer very slightly but it is important to note here that, due to limitations of the data already discussed and explored further later, the model does not profess to fit the data well. The highest R2 score it is capable of achieving in the model complexity graph is barely 0.3, only rising about 0.15 with increasing depths and then very slightly begins to diverge. Really any changes in max depth from 6 to 10 would not significantly impact the quality of the fit and it could never be accused of overfitting.

3. Creating the decision tree regressor and fit it with our training data - this was by far the easiest portion of the task. It worked totally out of the box. It seemed so easy that I assumed something must be wrong for a long time and was trying to debug it before I realised that it actually did just work out of the box like that. It seems too good to be true.

Simply put, you create the new Decision Tree regressor from sklearn with a parameter of "max_depth" set equal to the optimal max_depth decided by the previous step. You fit it with your target and feature training data.

Refinement

A great many things could be improved about this algorithm. I feel the project would benefit from an expanded range of features and more nuanced continuous variables which may be available as this data is supposed to be publicly available. My approach to making the variables continuous was a bit of a sledgehammer. Overall, I must concede that the project

was a bit of a failure in terms of predicting actual sentencing. I was optimistic about the possibility based on the fact that there was so much data but so many compromises and concessions had to be made to shoe-horn it into a regression problem when ultimately, a classification problem may have been much more fruitful. My personal interest in the topic, and likely my biases, blinded me to the limitations of this data.

IV. Results

Overfitting is a known issue when using Decision Trees however it's clear that this model is not being overfit as is has poor R2 scores on the training data. This doesn't mean that it will be able to generalise well. It does not generalise well at all.

I feel that this is most clear from the learning curves above. The performance of the model against training data deteriorates immediately as new data is added and then remains static at a low R2 score of about 0.3 from about 10,000 training points onwards. The score on the training data is so low after 10,000 records that the model could not be considered to be overfitting but it still has a low R2 score on the testing data. The ideal situation to be in is a learning curve where the testing and training data lines converge at a point with a high R2 score. This means it is accurate on the seen data but also that it has learned rules from the seen data and it is effectively applying it to the unseen data. For our model, this is not the case. It converges with a very poor R2 score very quickly on a small portion of the training data. If the algorithm is unable to adequately predict seen data, it is unlikely to cope well when given a testing set.

Model Evaluation and Validation

I chose a decision tree multiple regressor for this problem because, while I felt that the feature of weighted offence severity was very important and this was highly correlated with the sentence, I thought there might be a better fit for the data with increased dimensionality that was capable of representing some binary features such as sex and ethnicity. This theory, I feel, is born out by the results that with increased dimensionality, the model is capable of making ever so slightly more accurate predictions.

I feel that the methods selected and the implementation are proven to work based on the improvement on the benchmark models. However, it falls greatly below my expectations due to limitation in the data. It can generalise, but not well. It is hampered by a number of problems that are explained in detail elsewhere in the report.

Justification

The ultimate model performs better than the bench mark models but not significantly so. I

think by no stretch of the imagination can we consider this problem to have been solved but I am optimistic about the models capabilities if data was acquired with more nuanced continuous values for the duration of sentence, severity of the crime, age of the defendant and time taken to process.

The final model is more than 3 times better than the simple linear regression model. The decision tree regressors with varying max depths on the single feature of the most highly correlated feature returns values that are certainly an improvement on the simple linear regressor. The difference between the final model and the single decision tree regressor is not significant in my opinion.

R2 for the simple linear regression

0.10

Decision tree regressors on the most highly correlated features:

R^2 score for decision tree regressor with a max depth of 2 = 0.18

R^2 score for decision tree regressor with a max depth of 5 = 0.25

R^2 score for decision tree regressor with a max depth of 8 = 0.26

The result of the final decision tree regressor with optimised hyperparameters

R2 Score: 0.325214733155

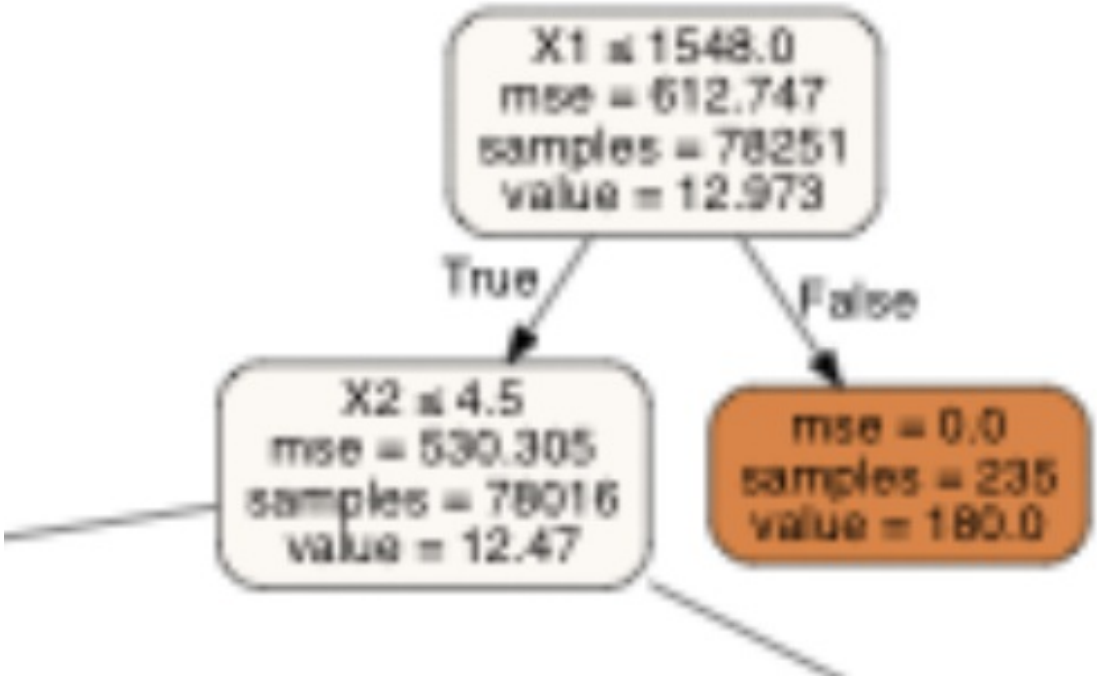
V. Conclusion

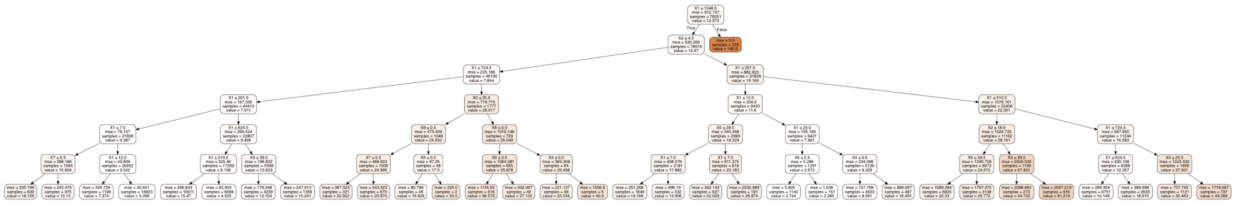
Free-Form Visualization

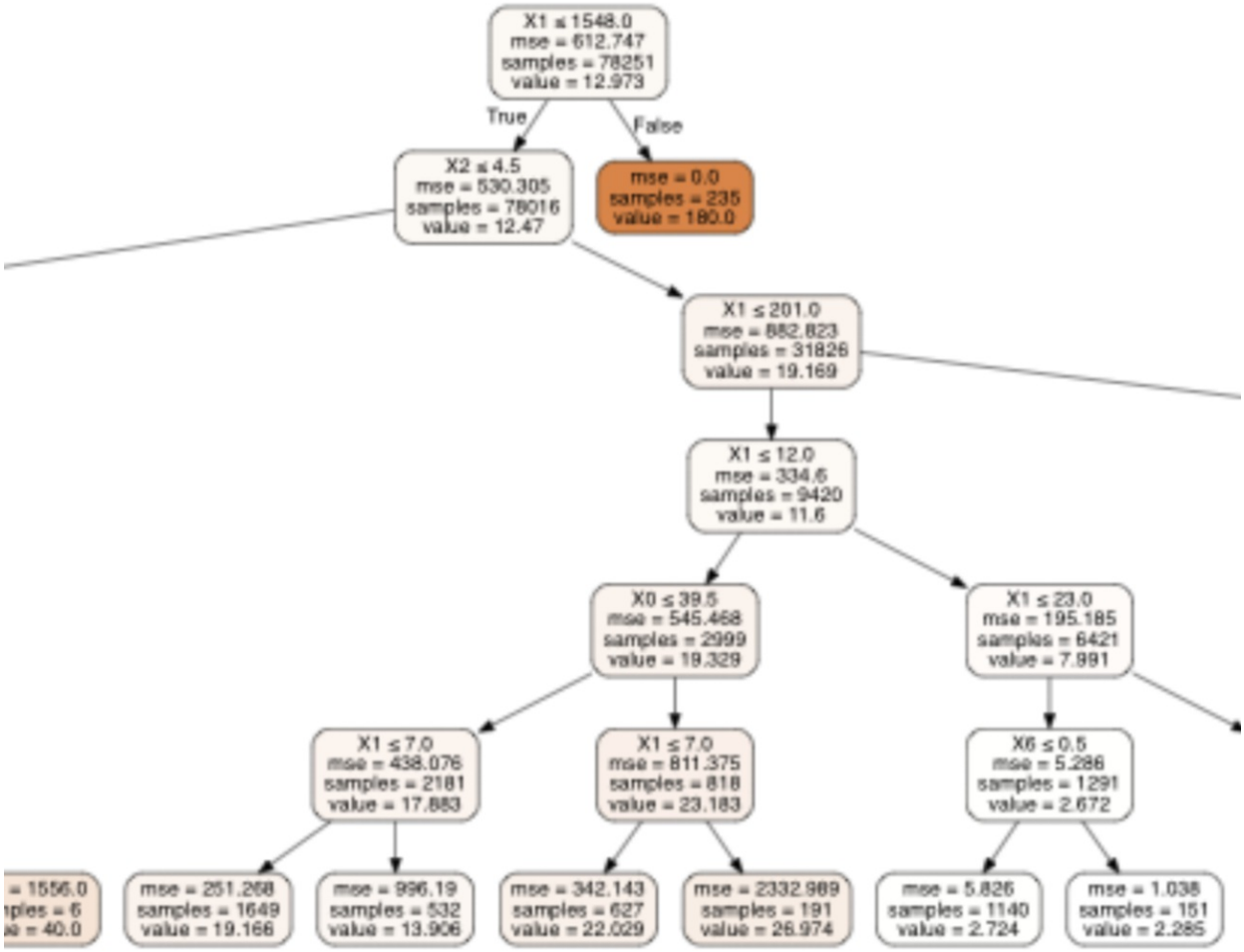
Below are the results of some the model's predictions on custodial sentences for different types of offences and different characterics of the defendants. I think fed with the correct data, this kind of result could potentially be very valuable to someone attempting to plan their life while involved in criminal proceedings.

I also include a visualisation of the final model as a decision tree. Here we can see the exact path taken by the algorithm in directing the samples to their predictions.

```
Defendant 1 is likely to be sentenced to: 49.29 months custodial sentence
Defendant 2 is likely to be sentenced to: 3.30 months custodial sentence
Defendant 3 is likely to be sentenced to: 8.56 months custodial sentence
```


X1 ≤ 1548.0
mse = 612.747
samples = 78251
value = 12.973
True
False
X2 ≤ 4.5
mse = 530.305
samples = 78016
value = 12.47
mse = 0.0
samples = 235
value = 180.0





Reflection

While I feel that I failed spectacularly in relation to my stated aim, I am pleased with my results as I fully comprehend now the specific issues with the data and my approach. I stumbled into a number of pitfalls from the get go.

It is hard to ignore the considerable issues in this dataset. I did not fully appreciate how limiting the total absence of nuanced continuous variables would be when I decided to undertake a regression based task. I feel the data is actually much more suited to a classification based task but I was too wrapped up in my desire to solve a very specific problem rather than solving the problems the data lends itself to. I am not particularly satisfied with my handling of the categorical variables through dictionaries but I think it is the best option given the data and the task. I think if I were to write the proposal again, I would focus on a classification task.

Having said that, some other sentencing data in a similar format that may at any time be released by the Ministry of Justice or could be requested as a Freedom of Information request. A more modern dataset may have more nuance. Additionally, some other supplementary data

could be very useful like the crime severity data tool from the Office of National Statistics.

While the final model does not meet my expectations of the solution, it does perform better than bench mark models and because of this, I consider it to be a success with caveats. My aim was to be capable of plugging in data to the model and receiving a ball park figure which, although likely wildly inaccurate, is provable better than a simple linear regressor and is a proof of concept for the idea that, given good data, such a model is possible.

Improvement

I cannot overstate the limitations I inflicted on myself by choosing data I was interested in rather than data that lends itself to such a problem. Most of my improvements would relate to not using dictionaries as a blunt instrument, not pressing ahead when the limitations of the data were obvious etc.

As said above, I think the use of dictionaries is unsatisfactory and for an improvement, I would like to seek out other datasets that might be supplementary to this one. Perhaps around the locations of the courts from large cities, the mean household income in the city of the court, the percentage of school finishers or NEET defendants in the city of the court (ie not in education, employment or training) or any other relevant datasets.

I'd also like to explore the problem from a classification point of view perhaps predicting the category of crime or otherwise.

Furthermore, I should have removed the outliers on reflection and narrowed the topic of the report to relate to offences excluding murder or other custodial life sentences.