Predicting types of crimes based on certain metrics

Group 1

Chloe Allan
ca14g14

Robert Anderson
ra12g14

Tom Barton
trb1g14

Jack Clarke
jnc1g14

Matt Cook
mc18g14

Nick Hobbs
nah1g14

Luke Pullman
ljwp1g14

ABSTRACT

This paper provides an insight into whether different types of crimes in England and Wales can be categorised using various data sets, such as weather, location, house price and population. To do this, public data was collected from data.police.uk, GOV.UK and the ONS and tested in combinations across different models in an offline supervised learning problem.

Different algorithms were explored including Decision Trees, K-Nearest Neighbours, Logistic Regression, AdaBoost and Random Forests. To compare results we looked at the accuracy of each algorithm primarily as well as precision and recall but found a classifier that always predicted one crime type was often more accurate than the models.

1 INTRODUCTION

The aim of this project was to identify unexpected patterns of crime in the UK. This would be done through the use of the open crime dataset that is released by the UK police, which is publicly available on their website [1].

The idea for this project was inspired by a Kaggle project that aimed to categorise crimes in San Francisco [5] based on the time the crime was reported and its location. In a similar manner, we took the UK data and framed it as a classification problem, predicting on the type of crime reported. This is a supervised problem that can be done offline so a wide range of techniques could be applied in this setting

With accurate predictions from the models produced, we would then be able to look into the most influencing variables for certain crimes to draw conclusions and patterns that we would not be able to do without the data.

We aimed to include more than just the initial dataset to influence our models in the hope that greater insights would out-perform our initial models. We thought information relating to current events, on a local or national scale, could also help us find patterns that influence which crimes are committed. Initial ideas around how we could collect this data were to use Twitter's public API for tweets that can be associated with the time and location of certain crimes.

Our project aims to produce a series of models that can make accurate predictions on our defined problem using this data. We will then look into the most influencing variables that these models use to influence their decisions, to find unexpected patterns in the data. We believe these techniques would be able to give us this understanding much quicker than looking at certain areas manually with a fine level of granularity to spot trends. From the start of this project we had the assumption that a model we produced could solve the problem we defined, whether that be with only the initial dataset or when more context is added through extra data.

2 DATA EXPLORATION

The initial dataset taken from the police website was created from a combination of the "street" data from every police force in England and Wales, this resulted in a CSV file with 16,701,247 rows. Table 1 shows the attributes available for this dataset.

Table 1: A table of the columns in the initial dataset, names and data-types. For greater detail, see [2]

Column Name	Data Type
Crime ID	String, unique alphanumeric
Month	String, in the form MM-YYYY
Reported by	String
Falls within	String
Longitude	Floating Point Number
Latitude	Floating Point Number
Location	String
LSOA code	String
LSOA name	String
Crime type	String[1]
Last outcome category	String
Context	String

As can be seen, a large number of these columns are strings that can hold a lot of meaning but are also difficult to process and input into many of the models we were proposing. Because of this and the fact that some of these columns were in fact empty for most of the rows, we decided to only keep the columns Month, Longitude, Latitude, LSOA code and Crime type.

Although the Month column is categorical, it can easily be transformed into two integer features, the month and the year of the crime. This will help our models find any seasonality in the data.

We also decided to keep Latitude and Longitude as it gives a very precise, floating point representation of the location of the

[1]The attribute we are trying to predict

crime. Which is far more useful to our models than what the Location column gave. Examples from that column were "On or near supermarket" or a named street.

LSOA codes are also useful for determining the location of the crime. LSOA, which stands for Lower Layer Super Output Area, are areas that are marked out to help report small area statistics in England and Wales therefore multiple LSOA codes can be within the same postcode [7]. Our initial dataset comprised of 34,749 LSOA codes in total for England and Wales (4 short of all LSOA codes possible).

Finally, we also kept the Crime type, the target attribute for our supervised problem. There are 14 types in total, which means we have a very varied multi-class classification problem. These categories are:

- Anti-social behaviour
- Bicycle theft
- Burglary
- Criminal damage and arson
- Drugs
- Other crime
- Other theft
- Possession of weapons
- Public order
- Robbery
- Shoplifting
- Theft from the person
- Vehicle crime
- Violence and sexual offences

Data cleaning was done before splitting train and test sets. This cleaning involved removing crimes that didn't have a location associated with them, either missing latitude/longitude or a LSOA code. The number of these empty location crimes removed was 753,170 which was 4.31% of the entire dataset.
Our training and test sets were produced from a stratified sample of our dataset using the `StratifiedShuffleSplit` in `sklearn`, splitting the data 80-20 for training and testing respectively.

2.1 Visualisation

With our dataset we can count the occurrences of crimes within the same area, whether that is the LSOA code or the postcode (discussed in section 3). This could then be used to spot any high-level patterns within our training set that we hope our models will also find. We would also be able to find any considerations that would influence the information that would affect the performance of our models.

Figure 1 tells you that the crime rates are very much related to the location (e.g., close to big cities such as Birmingham and Sheffield) which in turn can be related to population density.

It can also be seem that crimes categorised as "Drugs" appeared to be more common in areas with large shipping ports such as Liverpool, shown in Figure 2.

3 PRE-PROCESSING

In order to increase the performance of the classifiers, several pre-propcessing stages were added to attempt to clean some sections of the data which caused issues.

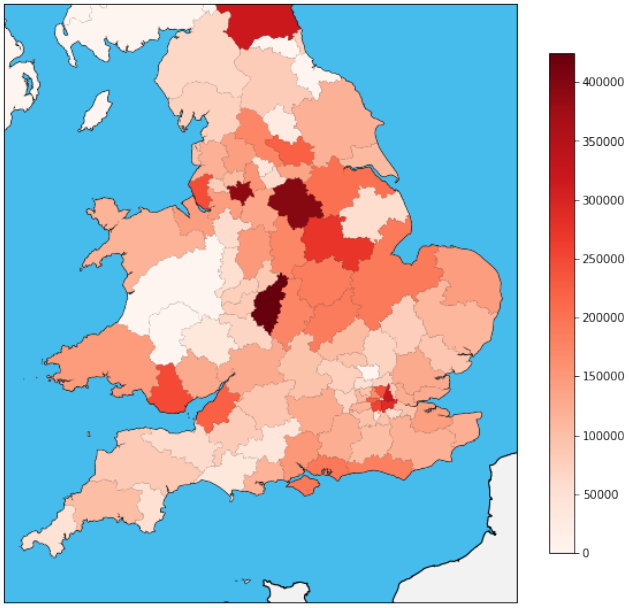

Figure 1: Map of all crimes by large postcode area

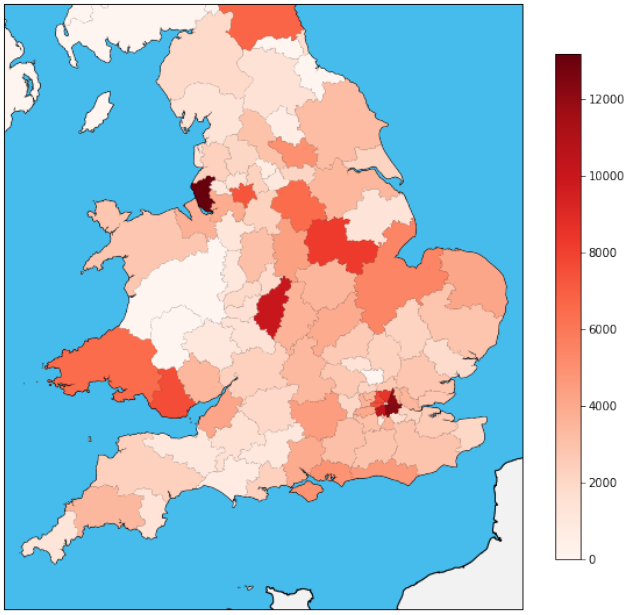

Figure 2: Map of crimes categorised as "Drugs"

As the month column contained information on both the month and the year of the reported crime, we could split this into two separate features. Using these, we would hopefully be able to find patterns depending on the time of the year that certain crimes take place.

The second operation was to scale each numeric value using the `StandardScaler` in `sklearn` which standardises features by

Table 2: Example postcode broken down into Area, District and Sector as well as the number of unique values for each.

Type	Postcode Part	Count
Area	SO	106
District	SO17	2215
Sector	SO17 1	7632

removing the mean and scaling to unit variance. We standardised our data to comparable scales to reduce the importance typically given by larger ranges in our analysis.

The dataset was also reduced to the exploration set which was 20% (2,672,248 rows) of the training set to speed up performance.

We found another dataset that gave us a mapping from LSOA codes to postcodes, meaning we could have greater insight into the area that the crimes took place in [8]. Postcodes can be broken down into area, district or sector, as shown in Table 2.

We used `sklearn's LabelBinarizer` to create one-hot vectors of non-numeric columns (such as the Area codes) so that they could be used within the models. This gives independent and orthogonal information about categorical data in a numerical format. We tried this with area code, of which there are 106, and this created a sparse matrix representation. This was far too much data for the algorithms to handle even though this was our smallest set of categorical attributes. We therefore did not use this data in the models that produced the results in the following sections.

4 INITIAL MODELS

Our initial approach was to create some quick models to see how they performed on the dataset.

We used location in the form of latitude and longitude alongside year and month to train our first classifier models. These models were compared with a base model that predicted "Anti-social behaviour" for every crime type as this was the most frequent. This base model outperformed all of the others, making it clear that additional data was required.

We used accuracy as our main measure for our predictive models and results from these can be seen in table 3.

Table 3: Models used as well as Mean and Standard Deviation of accuracy when run under a 3-fold cross-validation.

Model	Mean	Standard Deviation
ASBO	0.333099697	0.000677951
SGD	0.232224514	0.045177723
Logistic	0.333100071	1.96E-05
Perceptron	0.194424865	0.046641536
KNN	0.327180336	0.000328686
Decision Tree	0.26176762	0.000524516
Random Forest	0.287006294	0.000498665

`KNeighboursClassifer` from `sklearn.neighbors` was also added to the models. In order to set the number of neighbours for the classifier, several different iterations were performed with different numbers of neighbours to attempt to find the best, which came out at around 50 neighbours. `AdaBoost` ensemble classifier was also attempted, utilising the `DecisionTreeClassifier` with 600 estimators. The Final Classifier was the `RandomForestClassifier` from `sklearn.ensemble`, which used the default values set by `sklearn`. These classifiers were chosen as they were the most suitable for a multi-class classification problem, but also gave a range of different types of models, from Linear to Ensemble approaches.

4.1 Further Datasets

In order to develop a more accurate classifier, it was decided to add further data relating to the postcodes the crimes occured in. These datasets were then combined and added to the crime dataset to allow them to aid in classification.

4.1.1 Population. After initially attempting to use a dataset containing just a total population per LSOA code, a more detailed population dataset was sourced. This used data from the ONS to break down population by age and gender for each LSOA code, providing populations statistics for smaller geographical areas [9]. This data was tested with values for all ages from zero to 89 then 90+ as well as by grouping ages, with the gender split by ages. For example, `["Male age 0", "Male age 1", ..., "Male age 89", "Male age 90+", "Female age 0", ..., "Female age 90+"]`.

4.1.2 Affluence & property value. House Price data was also sourced to attempt to add metrics relating to affluency of a particular postcode. This sort of data is available from a number of sources, such as Zoopla and RightMove, however we decided to use a government provided dataset to avoid API restrictions [3]. The selected dataset included prices for property sales for the last 15 years and we adjusted these values for inflation using CPI [4].

To attempt to gain more data for each LSOA code, the Index of Multiple Deprivation dataset was added. Provided by the government, the dataset gives each LSOA code a score based on certain deprivation metrics, such as Income, Education and Employability, where higher scores indicate significant depravity [6].

4.2 Applying Models

This dataset was again unable to perform any meaningful classification, ASBOClassifier still being the best classifier on average, as shown in table 4. This seems to be due to erroneous data within certain postcodes which the classifiers cannot handle, so predicting just one class is generally better than actually applying meaningful models.

Table 4: Models used as well as Mean and Standard Deviation of accuracy when run under a 3-fold cross-validation.

Model	Mean	Standard Deviation
ASBO	0.34126	0.0007
SGD	0.245412	0.0514
Logistic	0.33617	0.0002154
Perceptron	0.175614	0.06425
KNN	0.32548	0.00096
Decision Tree	0.25644	0.0003654
Random Forest	0.312145	0.00010254

5 ADVANCED MODELS

As the initial models failed to outperform the ASBOClassifier, it was decided to add the datasets independently to attempt to identify which dataset had the largest affect on the performance, thus enabling the best accuracy to be achieved.

Also, due to the poor classification in the initial models it was decided to split the dataset up further to attempt to increase the accuracy, as a smaller region was being used for each model may allow patterns within that area to be identified. To do this, the data was split by area code, giving 106 different datasets to run the models against.

5.1 Affluency Data Models

By analysing crimes by type it is clear that "Theft from the person" occurs in wealthier areas, where the average house prices is significantly higher than for other crimes. However, this does not correspond with deprivation index scores, as shown in figure 3.

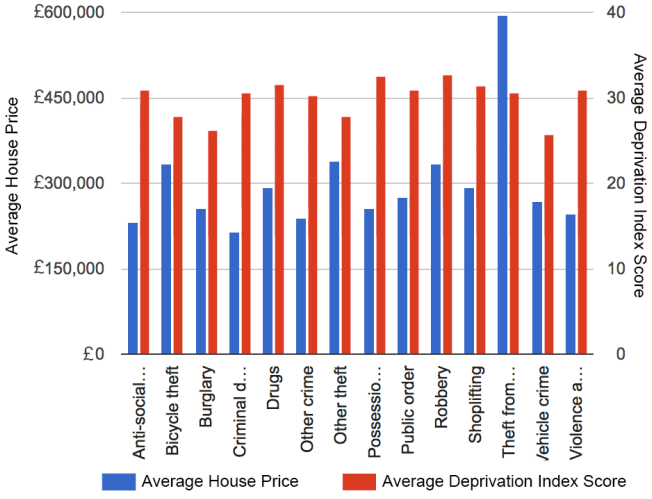

Figure 3: Chart showing average house price and deprivation index score for each crime type

To test this, the house price data was added to the dataset first. Doing this marginally increased the performance with 89 out of 106 of the models beating the corresponding areas `ASBOClassifier`. From the 89 that beat `ASBOClassifier`, 48 were from the `RandomForestClassifier`, 21 from `LogisticRegression` and 16 from `KNNClassifier`. The highest performance was from a `RandomForestClassifier` on area OR, which achieved an accuracy of 0.502.

The average accuracy for the `ASBOClassifer` was 0.3362, with the best classifier averaging 0.0225 greater than `ASBOClassifer`.

Table 5: House Price Model Results

Model	Mean	No. Best Performances
ASBOClassifier	0.33	18
LogisticRegression	0.33	21
KNNClassifier	0.34	19
RandomForestClassifer	0.34	44

Adding Deprivation Data gave similar results to the House Price dataset (accuracy increase of 2%), with minimal increase in accuracy across all areas. Interestingly, the ASBOClassifier dropped significantly, with KNN performing far better than previous iterations (42 best performances) while the RandomForestClassifier stayed around the same with 50 best performances. This suggests that the deprivation data may be more descriptive than the house price data, but is still unable to improve classification as the best model across all areas dropped to 0.452 accuracy.

Table 6: Deprivation Model Results

Model	Mean	No. Best Performances
ASBOClassifier	0.336	8
LogisticRegression	0.341	3
KNNClassifier	0.351	42
RandomForestClassifer	0.339	50

None of the above datasets have given a significant improvement to the models performance when compared to the basic models, with an average increase of 2%, with even the best performance only just reaching 50% accuracy (Area code OL). This ties in with the analysis, which suggested that there was no link between types of crimes and affluence of an area.

5.2 Population Data Models

To analyse this dataset, several visualisations were created to represent the total population for an area over the number of crimes within that area. Interestingly, as shown in Figure 4 the Salisbury postcode area is highlighted as the the worst for number of people per crime. To apply this to our problem, the total crime had to be filtered by a single type of crime to allow patterns to be identified.

The most interesting of these was the crime type Bicycle theft, which shows that there is significant hotspots in Plymouth, Salisbury and Huddersfield as seen in Figure 5. This suggests that there are hotspots for certain types of crimes when normalised by the population of that area.

The best dataset by far was the Population data, increasing the overall accuracy by 11.87% when comparing the best classifier for each area against the ASBOClassifer. However, the best accuracy still only just exceeded 50%, suggesting the accuracy increase has averaged out the accuracy across all the areas, as shown in Table 7 and Figure 7.

Table 7: Population Model Results

Model	Mean	No. Best Performances
ASBOClassifier	0.340	9
LogisticRegression	0.41	8
KNNClassifier	0.403	23
RandomForestClassifer	0.413	66

5.3 All Data Models

In an attempt to improve model performance, all the datasets were combined and split, then run against our models. This actually

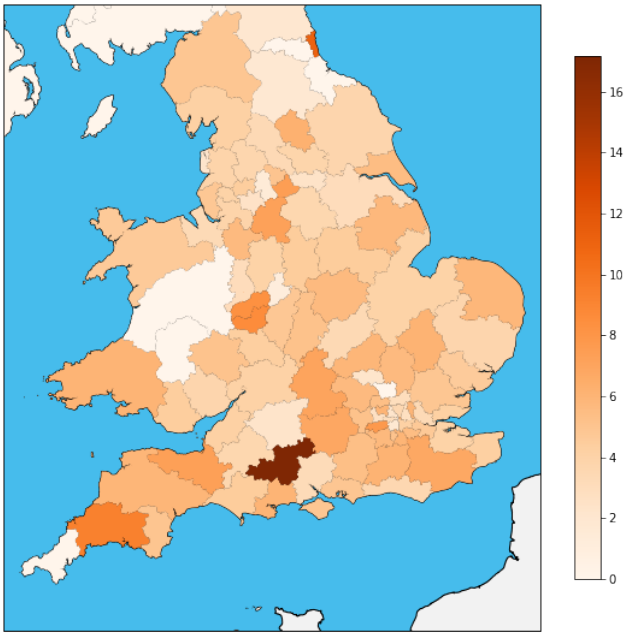

Figure 4: Map of all crimes by large postcode area

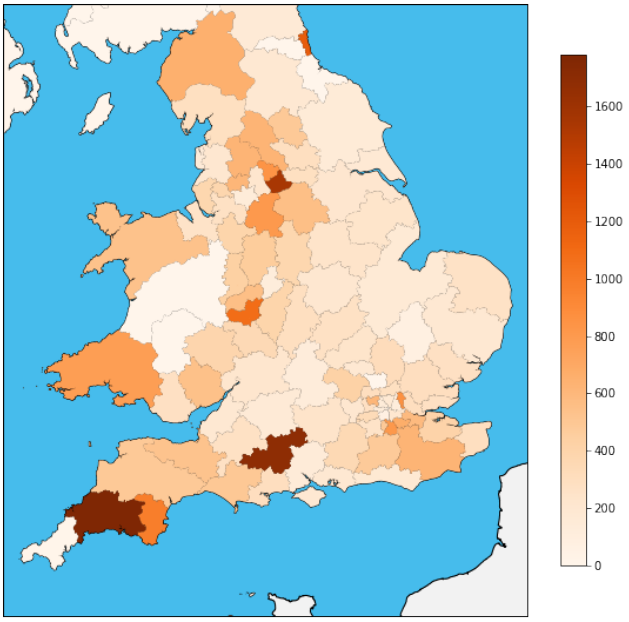

Figure 5: Map of crimes categorised as "Bicycle theft"

ended up reducing the accuracy compared to just the population dataset, with an accuracy of 7% when compared to the ASBOClassifier, shown in Table 8. This indicates that the house price and deprivation datasets have little descriptive power when modelling crime and are therefore removed from any further testing.

Figure 6: Chart showing the Accuracy for the Population and House Price datasets

Table 8: All Datasets Model Results

Model	Mean	No. Best Performances
ASBOClassifier	0.32	5
LogisticRegression	0.40	7
KNNClassifier	0.31	29
AdaBoostClassifer	0.34	0
RandomForestClassifer	0.41	65

5.4 Further Investigations

To investigate the further, additional performance metrics were added. These metrics were precision and recall, where precision indicates the proportion of positive identifications which were actually correct, while recall indicates the proportion of actual positives which were identified correctly.

When comparing precision and recall values from the basic data models to the current best models (Population dataset with RandomForest) it was found that across the crime types there was a significant increase in both precision and recall, with most notable changes being in the Shoplifiting category, shown in Figure 7.

Similar patterns were also identified within the precision results, with Shoplifting increasing from 0.31 to 0.53 when adding population data and Bicycle Theft increasing by 0.2 when compared to the base models. This indicates that the addition of the population data has allowed the models to improve accuracy for crime types which are more likely to be committed by younger offenders, such as Shoplifting.

6 REFLECTION & CONCLUSION

Overall this project was challenging as the models failed to provide greater insight into the data as they struggled to categorise the crimes with any accuracy. This is likely due to the lack of input data due to the anonymisation techniques implemented, as the date was vague (the San Francisco Kaggle project had times down to seconds) which could have helped draw further conclusions.

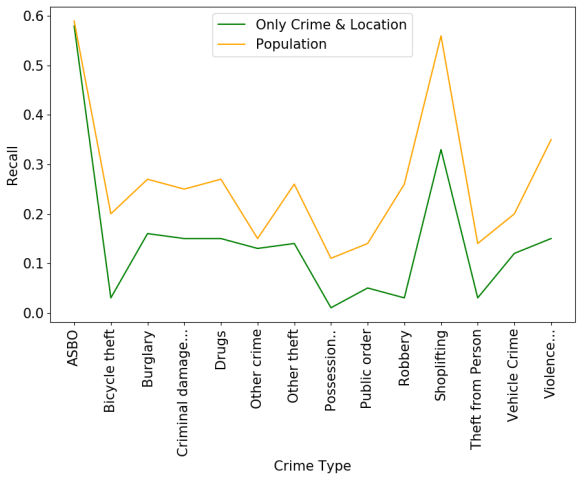

Figure 7: Chart showing the recall change between basic and advanced models

As well as this data-points were adjusted so that postal addresses had at least eight addresses, or none, and the centre points for roads were used as well as points of interest [1]. This means the crimes are not mapped to exactly the right location and therefore population data or affluence data may be incorrectly attributed to crimes. As well as this, having no data about the individual committing the crime severely limited the usefulness of the extra data we added, as it all related to the postcode rather than the individual themselves.

The crime data available covered England and Wales and so we focused on these countries. We had to drop social media because of the sheer amount of data as well as the format that it is present in, semantic analysis would have had to have been made which would have prolonged this project.

REFERENCES

[1] data.police.uk. 2018. Data downloads. (2018). https://data.police.uk/data/

[2] data.police.uk. 2018. Police Data, meaning of each column in the CSV file. (2018). https://data.police.uk/about/#columns

[3] HM Land Registry. 2014. HM Land Registry: Price Paid Data. (2014). https://www.gov.uk/government/collections/price-paid-data

[4] Inflation.eu. 2018. Historic harmonised inflation Great Britain HICP inflation Great Britain. (2018). http://www.inflation.eu/inflation-rates/great-britain/historic-inflation/hicp-inflation-great-britain.aspx

[5] Kaggle. 2016. San Francisco Crime Classification. (2016). https://www.kaggle.com/c/sf-crime

[6] Ministry of Housing, Communities & Local Government. 2015. English indices of deprivation 2015. (2015). https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015

[7] NHS.UK. [n. d.]. Data Dictionary Definitions: LSOA. ([n. d.]). https://www.datadictionary.nhs.uk/data_dictionary/nhs_business_definitions/l/lower_layer_super_output_area_de.asp?shownav=1

[8] Office for National Statistics. 2017. Postcode to Parish to Ward to Local Authority District (December 2011) Lookup in England and Wales. (2017). https://ons.maps.arcgis.com/home/item.html?id=c4aeb11ff5b045018b7340e807d645cb

[9] Office for National Statistics. 2017. Ward Level Mid-Year Population Estimates (Experimental Statistics). (2017). https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/wardlevelmidyearpopulationestimatesexperimental