Introduction to Programming

Master Program in Data Science
and Advanced Analytics 2018/19

Spatio-temporal analysis of crime rates and housing prices in London

Adeoluwa Akande[1], Carolina Araújo[2], Julian Kuypers[3] & Francisco Freitas[4]

[1] D20170353 - D20170353@novaims.unl.pt
[2] M20180262 - M20180262@novaims.unl.pt
[3] M20180409 - M20180409@novaims.unl.pt
[4] M20170062 - M20170062@novaims.unl.pt

Abstract

London has seen a surge in crime since 2016, with acid attacks becoming a common occurrence. Large metropolitan cities tend to have higher crime rates, and as such require highly efficient police forces to prevent criminal activity. This study aims to analyse street level criminal activity in London and average housing prices between 2016 and 2017 in order to determine if they are correlated. This may be important to disclose valuable information regarding the public good, and help the authorities optimize their resource allocation or to improve important services like the emergency call response. Using the free geospatial and scientific packages available in Python, areas with higher criminal activity were identified, and correspond to the areas with the highest average housing prices. These findings lead us to believe that both events are correlated. Unexpectedly, our analysis revealed that the high-crime, high-price areas identified correspond to the city's busiest tourist spots, which leads us to believe there is a relation between tourism and crime. Although this makes sense intuitively, further research would be needed to prove this. We conclude our study with the implementation of a few classification algorithms available in Scikit Learn, and provide overview of results for each model.

Keywords: spatial exploration; exploratory data analysis, predictive models; python programming, crime analysis.

Statement of Contribution: Conceptualization - Adeoluwa Akande, Carolina Araújo, Julian Kuypers & Francisco Freitas; Aspatial Data Cleaning - Julian Kuypers; Spatial Data Cleaning - Adeoluwa Akande; Exploratory Analysis - Adeoluwa Akande, Carolina Araújo & Julian Kuypers; Spatial Analysis - Adeoluwa Akande & Francisco Freitas; Modelling - Julian Kuyper; Writing (review & editing) - Adeoluwa Akande, Carolina Araújo, Julian Kuypers & Francisco Freitas

Introduction to Programming

Master Program in Data Science
and Advanced Analytics 2018/19

I. Introduction

This report gives an overview of the data used to perform our study of London street-level crimes and housing prices between 2016 and 2017. After the dataset has been described, extraction and cleaning methods are explained; here we provide insights on the logic used to fit the data for our purposes. The analysis section of this reports presents our spatio-temporal exploration of the data, as well as the insights gained from our analysis. Finally, an brief overview of the different classification and prediction models used is given, along with the results obtained and some insights on how to improve them. Our conclusion aggregates our findings to confirm our hypothesis; crime rates in London are influenced by housing prices.

II. Data

a. Description

Two datasets have been retrieved. Both datasets are available at Data Police UK (https://data.police.uk/data/). The data on this site is published by the Home Office, and is provided by the 43 geographic police forces in England and Wales, the British Transport Police, the Police Service of Northern Ireland, and the Ministry of Justice.

A third data set containing the average housing prices in London Boroughs has been used. The Land Registry (https://data.london.gov.uk/dataset/average-house-prices), a non-ministerial department, registers the ownership of land and property in England and Wales. They publish the annual mean and median property prices calculated by the Greater London Authority/City Hall. The data has been aggregated to Borough, Ward, MSOA, LSOA, Port Code, Postcode District, and Postcode Sectors.

To explore changes throughout the time, spatial information can represent a valuable asset. Datasets containing geographic location (i.e. coordinates or geographical areas) have been selected. Specifically, shapefiles of the administrative units of London was gotten from (https://gadm.org). Python includes some options to tackle geographical features. One of the Python's derived tools is Folium, an engine for building interactive maps. Folium makes it easy to visualize data that's been manipulated in Python on an interactive Leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing Vincent/Vega visualizations as markers on the map. The library has a number of built-in tilesets from OpenStreetMap, MapQuest Open, MapQuest Open Aerial, Mapbox, and Stamen, and supports custom tilesets with Mapbox or Cloudmade API keys. Folium supports both GeoJSON and TopoJSON overlays, as well as the binding of data to those overlays to create choropleth maps with color-brewer color schemes[1,2]. Given the amount of the data points available, by using Leaflet maps it was possible to overcome rendering limitations of GIS offline packages.

Metropolitan Police Services/City of London Police Jan 2016 - Sep 2018:

The UK Police street-level crime datasets contain the following features:

- ***Crime ID*** (str): A unique identifier for each crime.
- ***Month*** (Datetime): The month and year the crime was registered
- ***Reported by*** (str): The police instance that filed the report - can be either the City Police which operates in the 'City of London', or the Metropolitan Police, which operates in all Boroughs (Greater london).
- ***Falls Within*** (str): Which jurisdiction the crime falls in; City Police or Metropolitan Police
- ***Longitude*** (float): Anonymised X coordinate on the map where the crime occured
- ***Latitude*** (float): Anonymised Y coordinate on the map where the crime occured
- ***Location*** (str): The name of the street where the crime occured, or a generic description of the crime location (e.g nightclub, petrol station, tube station. etc.)
- ***LSOA Code*** (str): Lower Layer Super Output Areas identifiers, a geographic hierarchy designed to improve the reporting of small area statistics in England and Wales.
- ***LSOA Name*** (str): Names of Boroughs, according to the Lower Layer Super Output Areas.
- ***Crime Type*** (str): Broader categories describing the crime
- ***Last Outcome Category*** (str): Outcome of the arrest or interpellation
- ***Context*** (str): A field for police forces to provide additional human-readable data about individual crimes.

The Average House Prices - Boroughs 1995 - 2018:

This data holds in information regarding average housing prices in the different Borough of London between 1995 and today and contains the following features:

- ***Code*** (str): Lower Layer Super Output Areas identifiers, a geographic hierarchy designed to improve the reporting of small area statistics in England and Wales.
- ***Area*** (str): Names of Boroughs, according to the Lower Layer Super Output Areas.
- ***Year Ending month year*** (float): Contains the mean housing price in the area calculated at a given period (quarterly).

Geospatial Information

Crime data tables are georeferenced - all data points include longitude and latitude values in decimal degrees. Basemaps were retrieved from OSM. The projection system is WGS84 Web Mercator.

b. Extraction

Our housing data came in an excel sheet which contained several sheets; mean housing price, median housing price, sales, and Lower Quartile values. We extracted all of them but only used the mean housing prices.
Our crime data came in csv format, with each month as a separate file. All files needed to be concatenated in order to get all data the into a single file.

c. Transformation

London Housing Data

The Average Housing price was narrowed to only the Boroughs within the Greater London area, and we only kept the data from 2014 to 2017 as we believed that this period would give us insights as to if and how changes in price can affect crime rates. Our dataset reported quarterly average prices, and we needed a monthly price to match to our crime dataset. In order to achieve this we created new columns for the missing months and created an estimated average price for these months by dividing the difference between two consecutive known months by two, so as to get an average increase or decrease. This value was plugged into the respective months that we created.

The dataset was transformed so that each row would hold an area (Borough), Month, Year, and Average Price; this made our dataset more readable and would make it easier to later merger with our crime data.

London Crime Data

The City of London and Greater London datasets hold the same columns and value types, so it was just a matter of merging both datasets into one (LondonStreet) and cleaning the data. We decided to work with data from 2016 and 2017, as those were the only two complete years of data.

Merged Data

Our crime (LondonStreet) and housing (housingMean) datasets were merged on Month, Year, and Borough. This resulted in a dataset (StreetCrimePrice) which contained all crimes between 2016 and 2017 as well as the the current average housing price in the area of the crime at the time. Our final model used for analysis is called FinalModel, and only holds the columns relevant to our analysis and exploration.

III. Results and Discussion

Analysis

Real estate in London has seen a steep increase in value since the late 1990’s (ONS, 2014). Our aim is to study the potential effect this increase has had on crime rates. In order to study the relation between housing prices and crime, we need to understand which are the most prominent crimes, which are the Boroughs that are most affected by crime, and how crimes are distributed throughout the year.

1. Crime Analysis

To reach our goal, first we had to do some exploration analysis regarding crime. The crime rate between 2016 and 2017, although it was high it did not vary that much (rounding 1.000.000 number of crimes). Analyzing this timeframe, crimes vary in winter and summer. We understand that the peak of crime is in summer (July), reaching around 65.000 crime and drops in winter reaching its lowest value of about 5000. Although the number of crimes are very similar in both years, we can see that the year 2017 was a bit worst, specially in March.

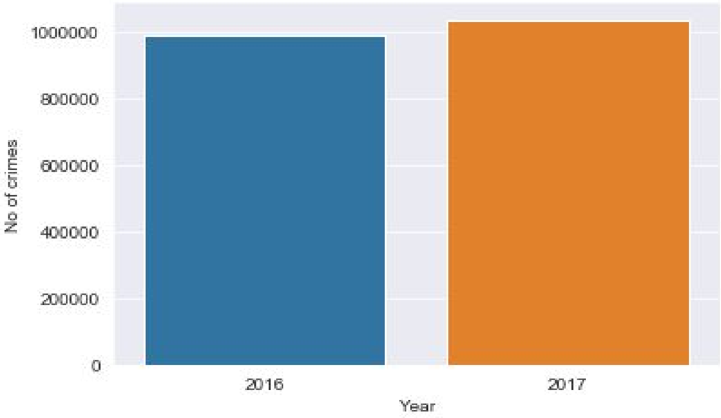


Figure 1: Number of crimes in London between 2016 and 2017.

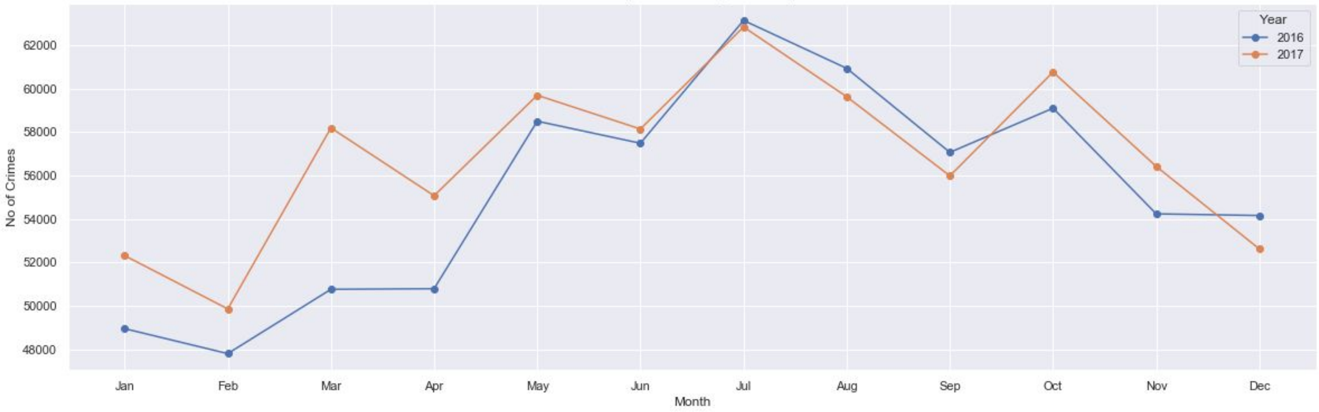


Figure 2: Monthly evolution of number of crimes in London (2016-2017).

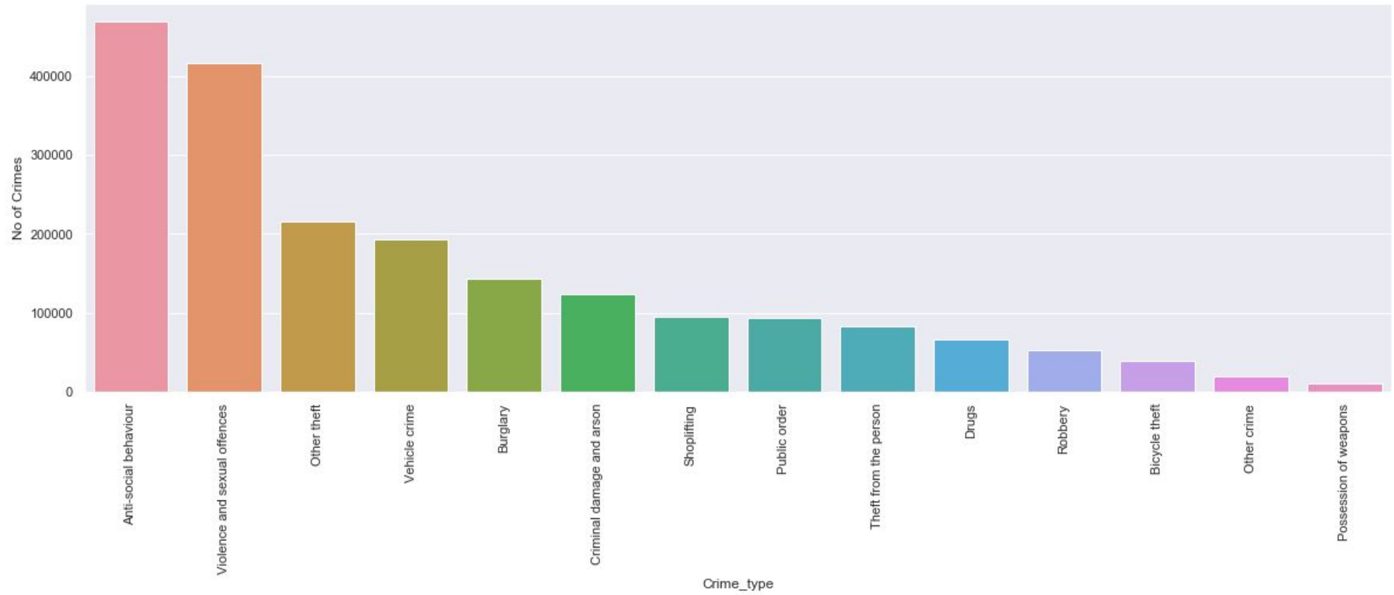
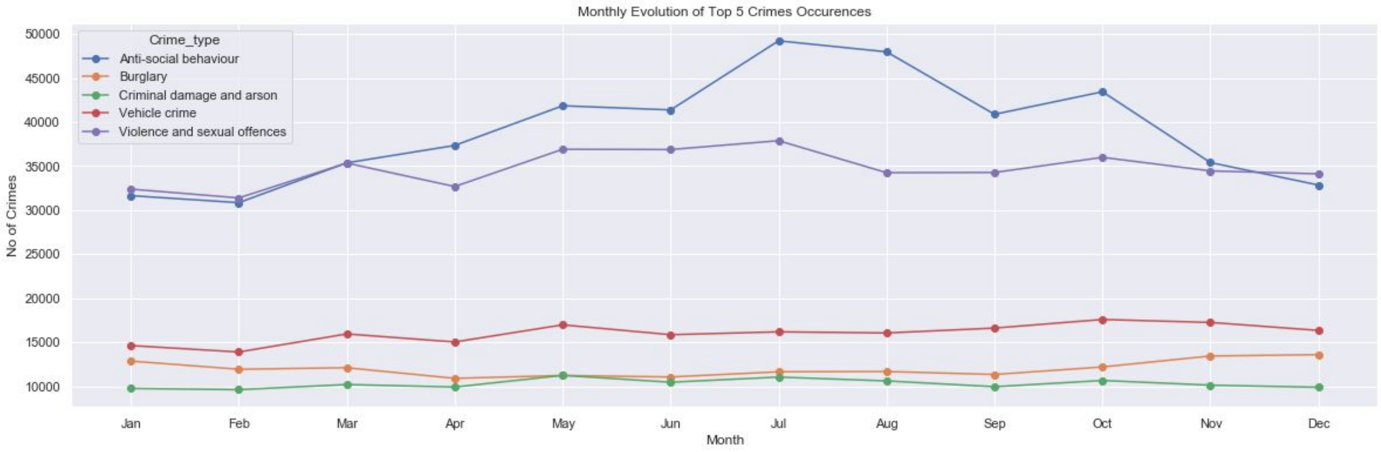


Figure 3: The most committed crimes in London (2016-2017).

After plotting the distribution of crime types, we understood that certain crimes were more prominent than others. Based on this information, we decided to do further analysis in crimes that occurred more than 100.000 times; this would narrow down our crime types to the five most important ones: **Anti Social Behaviour**, **Violence and Sexual Offenses**, **Vehicle Crimes**, **Burglary** and **Criminal Damage and Arson**. Analyzing these top crime types by month, we can see that the anti social behaviour and violence and sexual offences behave according as the graph below; they peak July and October. The rest of the crime types have a constant behaviour throughout the year.

Figure 4: Monthly evolution of the top 5 crimes in London (2016-2017).



We selected the top 10 boroughs that have more crime, concluding that Westminster has a significantly high crime rate (135.000 crimes) compared with the other nine boroughs. As seen from the heat map below, the top 10 boroughs have more correlation with criminal damage and arson and burglary, meaning that these are the type of crimes most common in the boroughs.

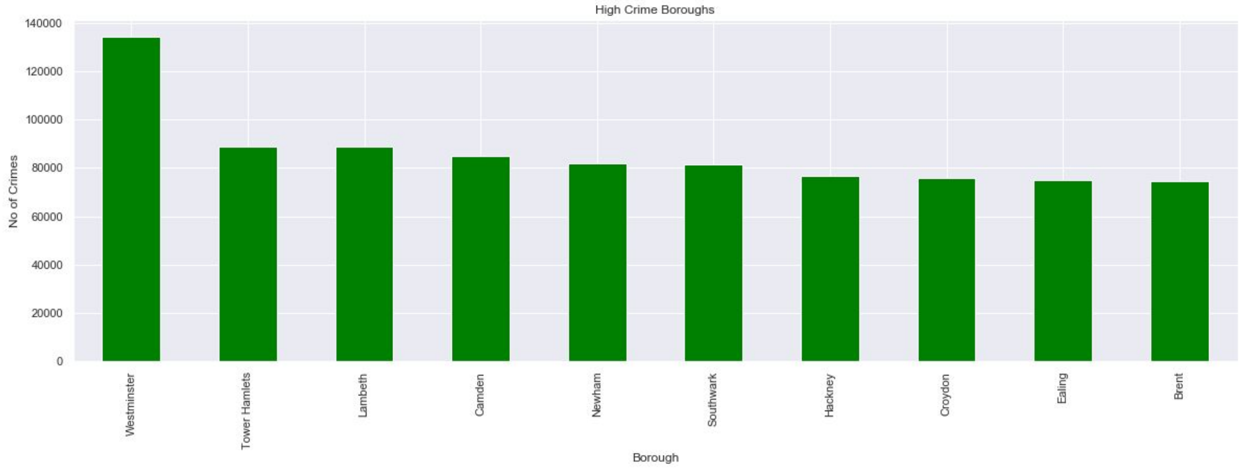


Figure 5: London boroughs crime rate (2016-2017).

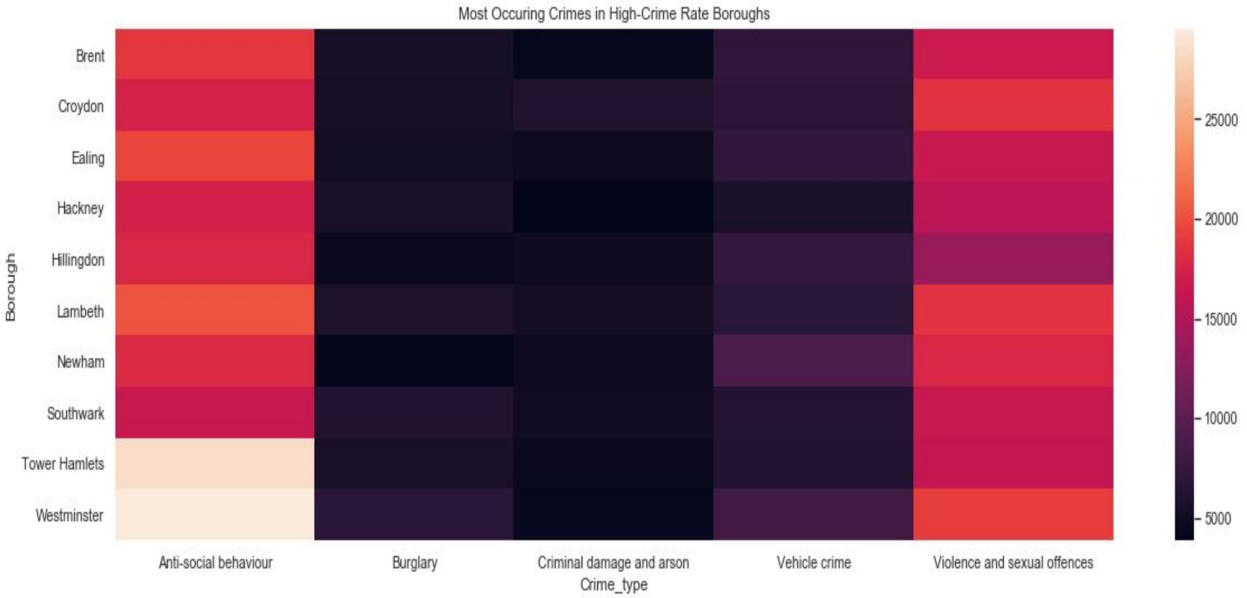


Figure 6: Heat map of the most occurring crimes in top 5 crimes in London (2016-2017).

Below we can see that the number of crimes with status finished with no suspect founded and status update unavailable are too high compared with the other status, around 750.000 and 600.000 respectively.

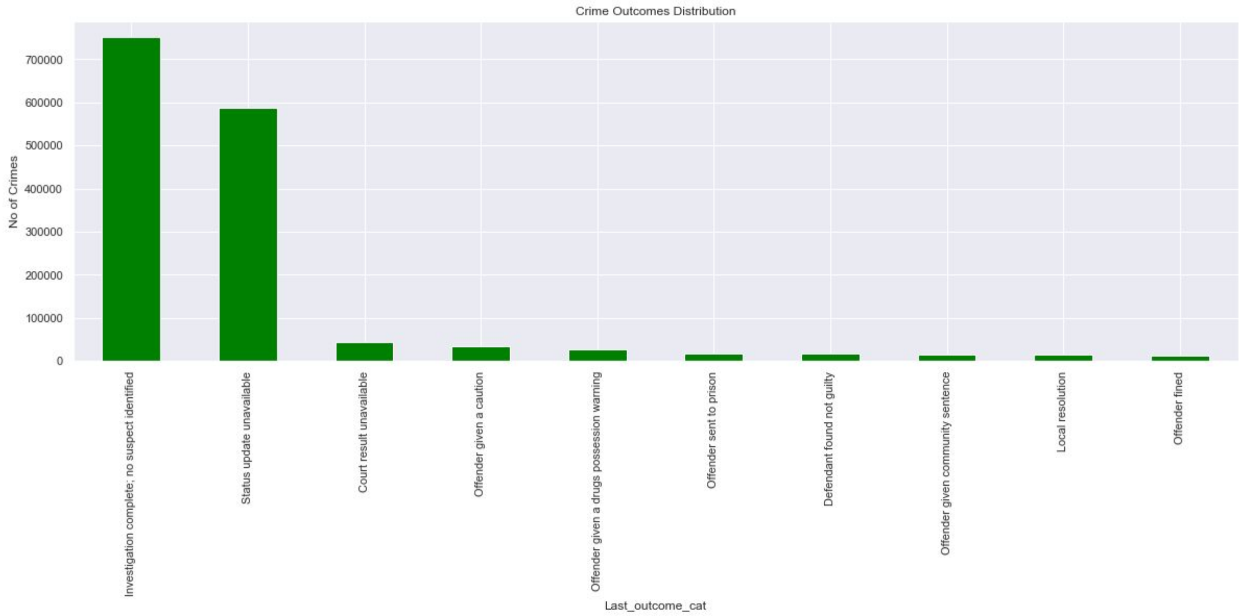


Figure 7: Crime outcomes distribution (2016-2017).

2. London Housing Prices

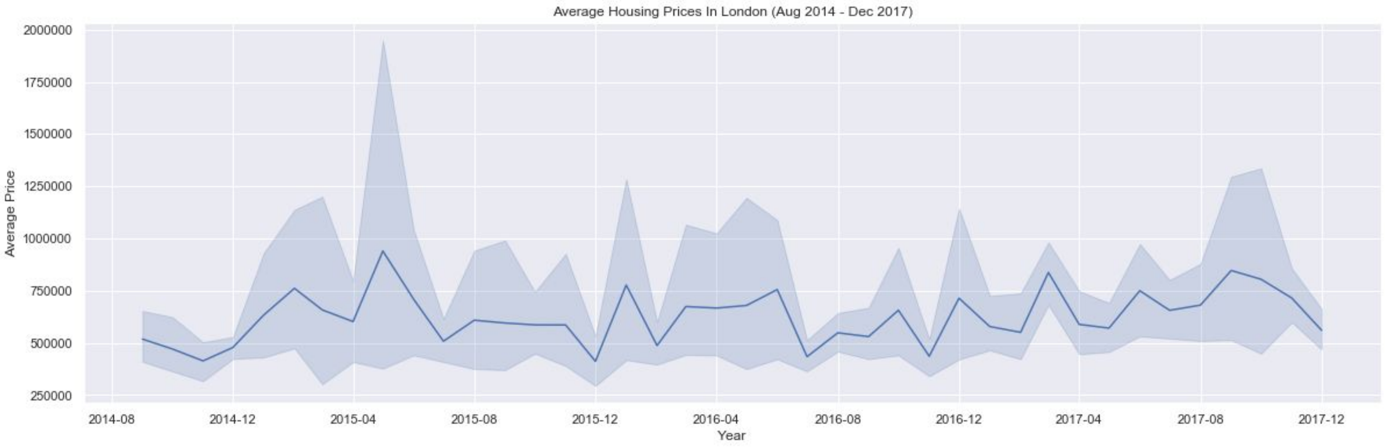


Figure 8:Average housing prices in London (2014-2017).

While the average housing price has seen peaks and bottoms between 2014 and 2017, we can clearly see that there are high deviations from the mean. This variance seems particularly high during certain months; December and August seem to be the periods where both average prices and highest prices increase, while October/November seems to be the times when the market is at its lowest.

When studying the Borough-level average housing prices, we notice that there are a few central and West London Boroughs that have average prices up to four times higher than the other Boroughs. Out of the five Boroughs with the highest average housing prices, four are tourist attractions: Kensington and Chelsea for their luxurious boutiques, Westminster for its monuments and historical sites, Camden for it's alternative market, and the City of London for its restaurants and bars.

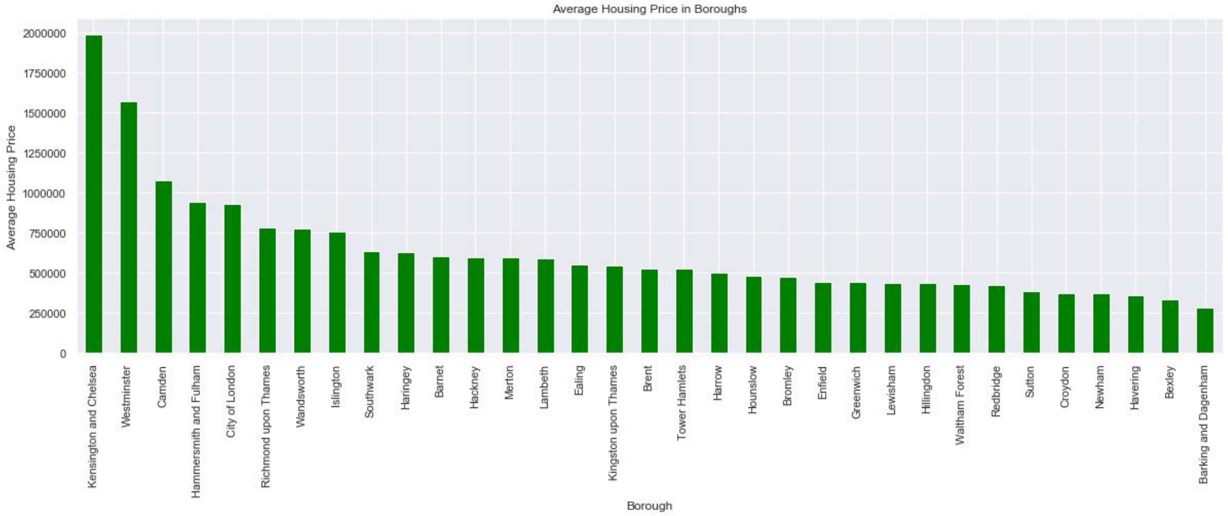


Figure 9:Average housing prices in boroughs of London (2016-2017).

Spatial analysis

A spatial analysis was made to have a more clear understanding of crime and housing prices in the boroughs of London.

1. Housing Prices

We can deduct from the graph below that the housing prices in London are higher in the boroughs Westminster and Kensington and Chelsea.

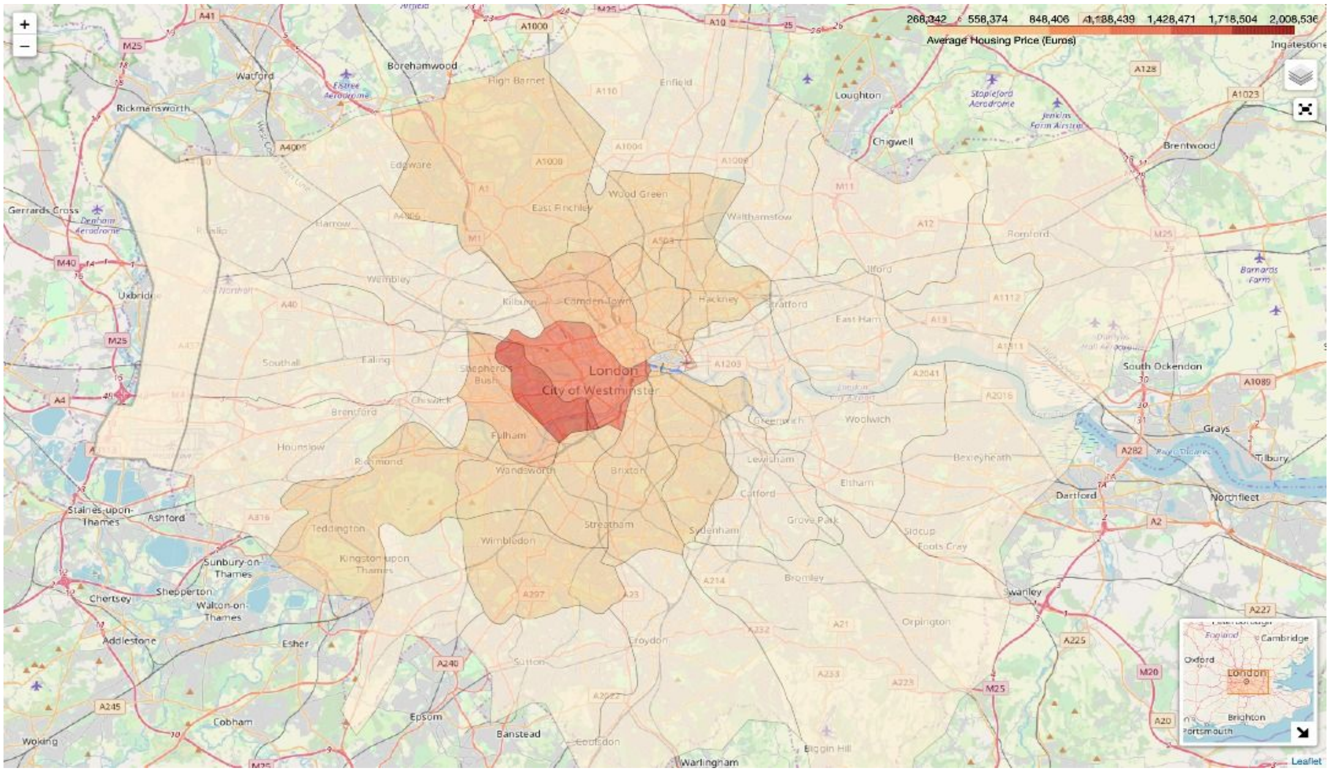


Figure 10:Choropleth Map - Housing Prices in London.

2. Crime

The graph below clearly shows that the number of crimes are higher in the center of London, with 169.487 of crimes, rather in the rest part of the city.

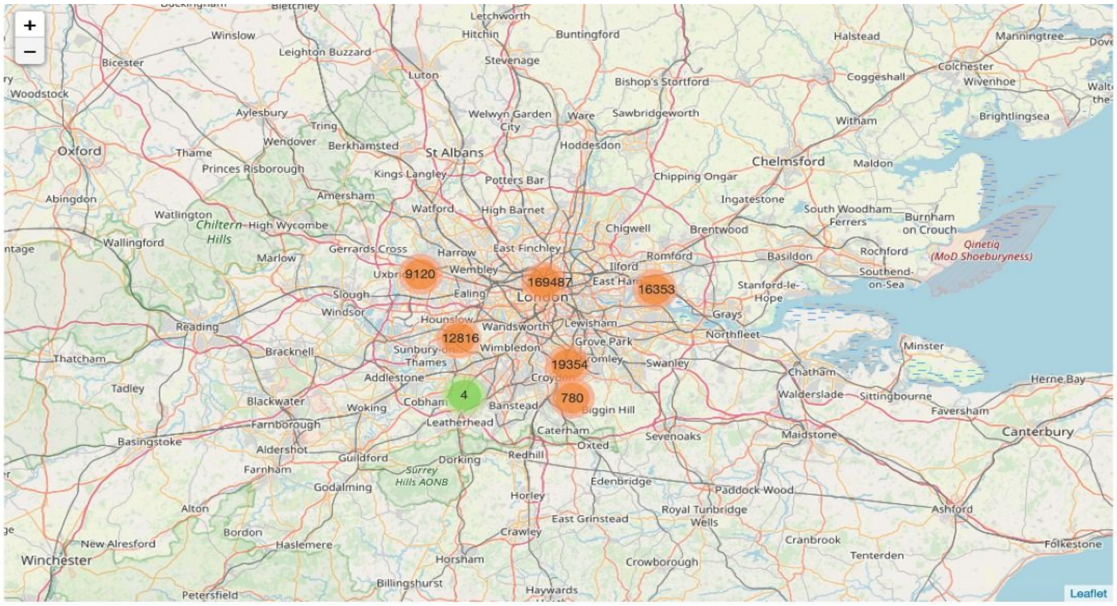


Figure 11: Clusters Plot - Crime Counts in London.

When we take a closer look at the boroughs of London we can see that Camden, Westminster and Kensington and Chelsea, all together, have the highest number of crimes (70.360).

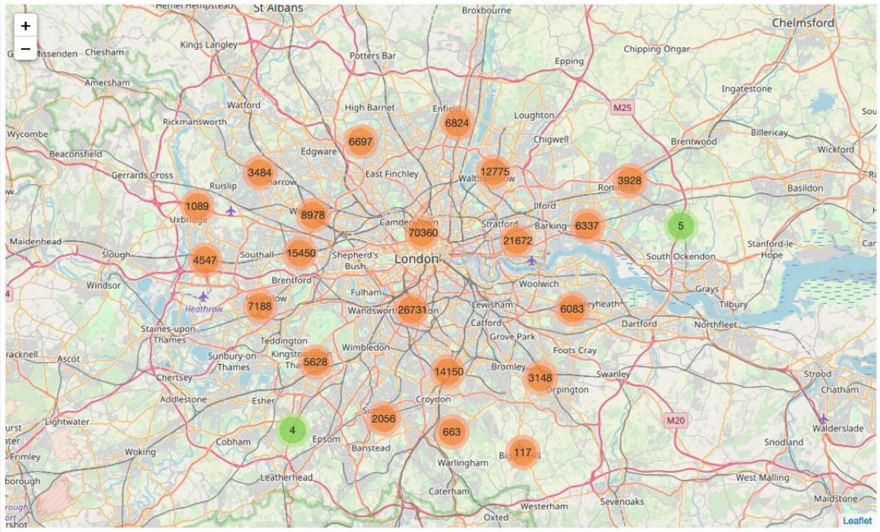


Figure 12: Clusters Plot - Crime Counts in boroughs of London.

This is the choropleth map of the counts of crimes in London. Comparing it with the choropleth map of the housing prices in London, we can see that both maps roughly follow the same spatial pattern; high in the center and decreases as we move to the outskirts.

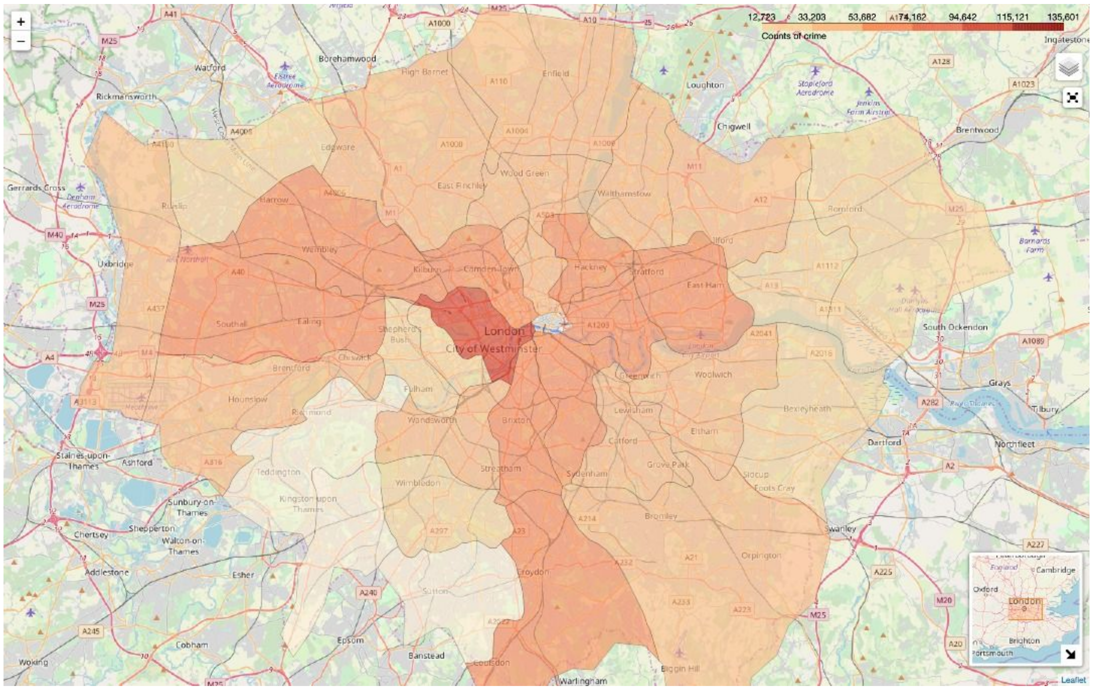


Figure 13: Choropleth Map - Crimes in London Count.

Prediction Models

Under Data Science, a model consists of an attempt to understand and represent reality through a particular lens. A model is therefore an artificial construction where all the extraneous details are removed or abstracted. The abstracted details must be analysed in all the details to check what might have been overlooked by the model [3]**.**

Classification

In this section, we will give an overview of the different classification models used to analyse and predict crimes in London; the target feature is crime types, and the model features are Longitude, Latitude, Location, Reported_by, Month, Year, Borough, and Price. Since we are studying the effect of housing price on crime rates, we will compare the models' performance where the average housing price is included, and where it is not. To fit our models, our categorical data (Borough, Location, Reported_by, crime type) were encoded using Scikit Learn's Label Encoder[4].

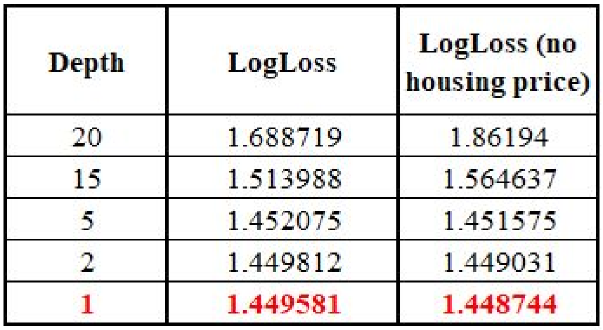
Performance Evaluation

Logarithmic loss (related to cross-entropy) measures the performance of a classification model where the prediction input is a probability value between 0 and 1. The goal of our machine learning models is to minimize this value. A perfect model would have a log loss of 0.

$$logloss = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log(p_{ij})$$

1. Random Forest Classifier

The random forest classifier creates a collection of decision trees, where it randomly selects features and observations, and averages the results. We established that the best number of trees is around 60 (Appendix 2).With regards to the quality of the split, the Gini impurity function performs betters, although the difference is minimal. We were surprised to find that a lower depth performs best, as our intuition was that the deeper the tree, the more precise it would become when making decisions.



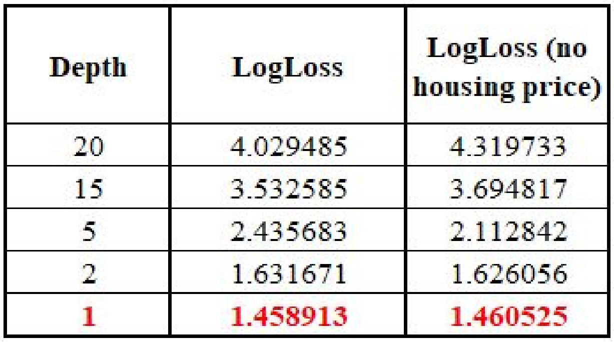
Depth	LogLoss	LogLoss (no housing price)
20	1.688719	1.86194
15	1.513988	1.564637
5	1.452075	1.451575
2	1.449812	1.449031
1	**1.449581**	**1.448744**

Figure 14: Random Forest Classifier in crime with and without housing prices.

2. Gradient Boosting Tree Classifier

We extend our random forest classifier with gradient boosting, a sequential algorithm where predictors learn from the mistakes of the previous predictors; in theory this should help us reduce variance.

A new parameter is introduced; the learning rate which controls the rate of adjustment of the weights with respect to gradient loss (i.e the difference between the predicted output and the actual output). Our analysis shows that a lower learning rate, which leads to smaller corrections in the model, perform best (Appendix 3.).



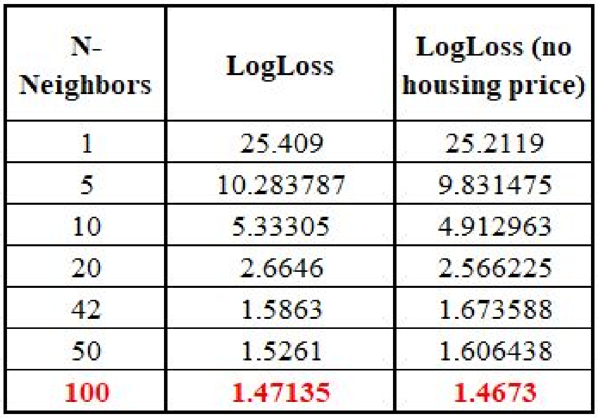
Depth	LogLoss	LogLoss (no housing price)
20	4.029485	4.319733
15	3.532585	3.694817
5	2.435683	2.112842
2	1.631671	1.626056
1	**1.458913**	**1.460525**

Figure 15: Gradient Boosting Tree Classifier in crime with and without housing prices.

Our results were slightly worse than with the Random Forest Classifier, although the difference is negligible. It is possible that our model was overfitting, which is a risk when running a boosting algorithm for too long.

3. K-Nearest Neighbor

Another powerful classification model we wanted to test is the KNN; an unsupervised, lazy machine learning algorithm that uses a distance function to measure feature similarity. It determines how classes are separated to predict the classification of new data.



N-Neighbors	LogLoss	LogLoss (no housing price)
1	25.409	25.2119
5	10.283787	9.831475
10	5.33305	4.912963
20	2.6646	2.566225
42	1.5863	1.673588
50	1.5261	1.606438
100	**1.47135**	**1.4673**

Figure 16: KNN in crime with and without housing prices.

We tested our KNN model by changing the weights, distance metrics, and algorithms, but the performance was the same regardless which methods were used (Appendix 4.). Only the number of neighbors affected the performance of our model.

4. Logistic Regression

The underlying technique of a logistic regression is the same as a linear regression, with added functions that allow us to use it for classification problems instead of only continuous dependent variables.

The best LogLoss was obtained using the solver Limited-Memory BFGS method, although the difference are negligible. Our best LogLoss is 1.4497.

5. Principal Component Analysis (PCA)

In order to improve the accuracy of our models and reduce overfitting, we can reduce dimensionality by using PCA, which finds a new set of dimensions that are linearly independent and ranked according to the variance of data among them. We found that our models would probably perform better with 3 components.

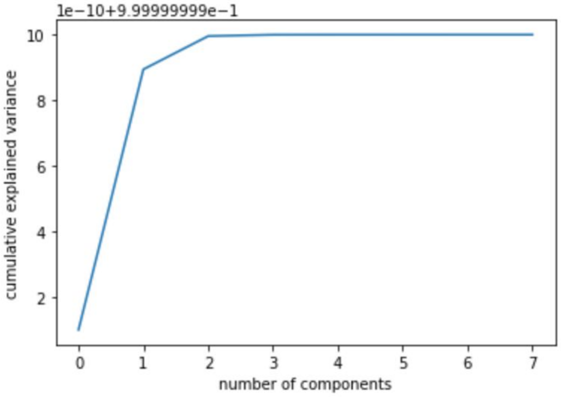


Figure 17: PCA for ideal number of components.

The Random forest classifier seems to be the best model for our purposes. To further improve our model, we can fine tune the parameters even further. We could also explore other encoding methods such as one hot encoding, which is a representation of categorical variables as binary vectors. This way our variables couldn't be interpreted as having hierarchical meaning due to the number assigned to them during encoding.

IV. Conclusions

After analyzing housing prices and street crimes, we can conclude that there is a correlation between both variables; areas with high average prices tend to have higher crime rates. Both Westminster and Camden have some of the city's highest housing prices and crime rates. While we can not confirm that one happens because of the other, our analysis shows that they bound in some way. With the observed, we could argue that Boroughs with high tourist concentrations are attractive to criminals; Westminster, Camden, Tower Hamlets, and Southwark are home to some of London's most iconic tourist attractions, and have the city's highest crime rates, particularly in anti-social behaviour. However this is only a hypothesis, and would need further study to prove the correlation.

V. References

[1] R. Story, *Folium: Make Beautiful Maps with Leaflet.Js & Python* (2015).

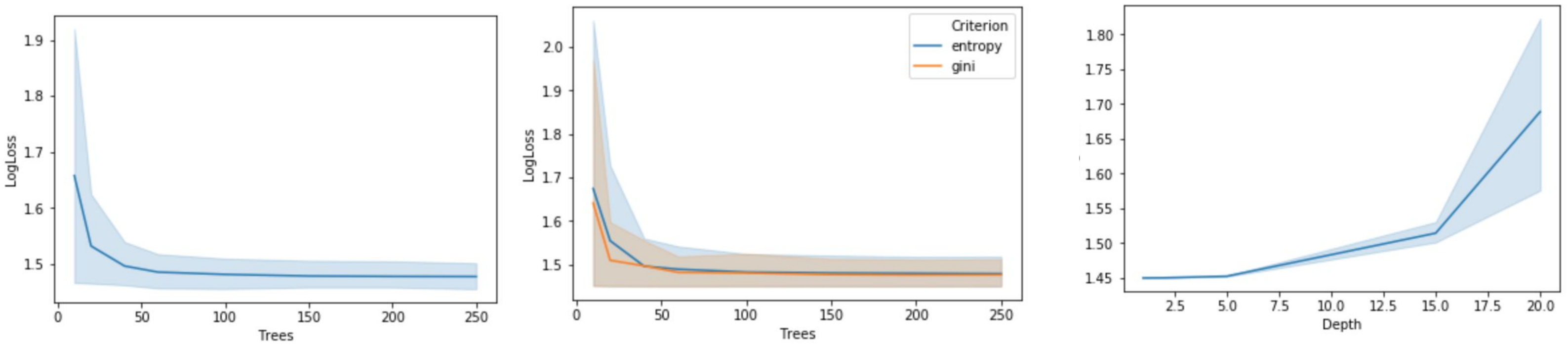
[2] R. Story, (2013).

[3] R. Schutt and C. O'Neil, *Doing Data Science: Straight Talk from the Frontline* (O'Reilly Media, Sebastopol, 2014).

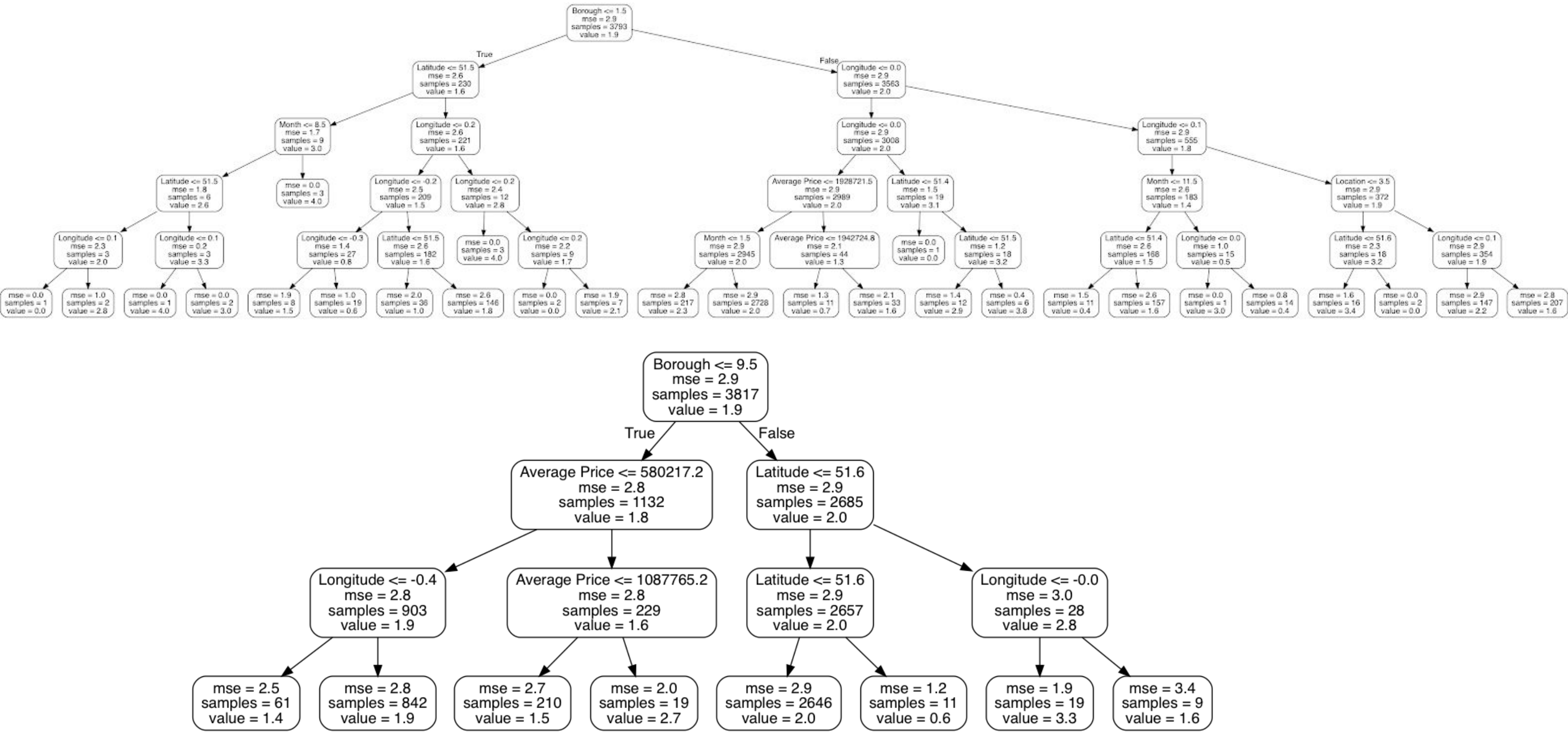
[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res. 12, 2825 (2011).

VI. Appendix

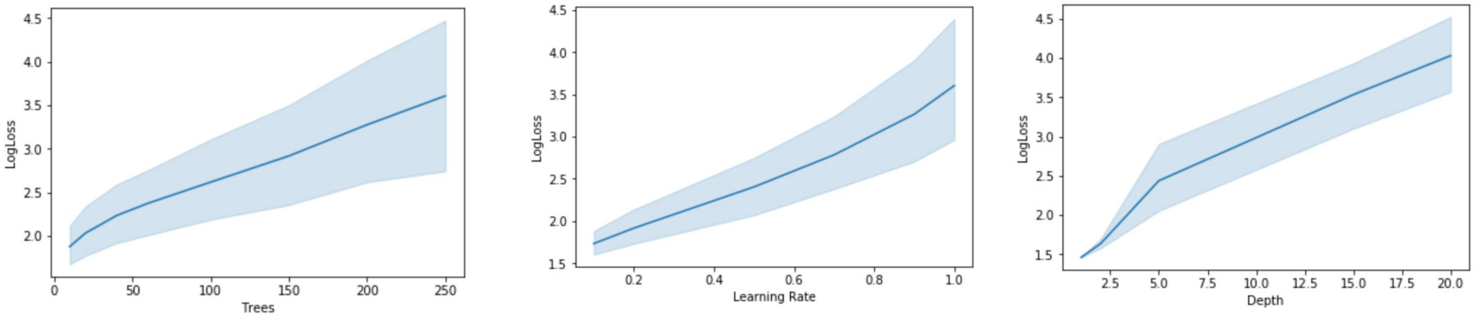
Appendix 1. Average LogLoss for Random Forest Classifier - Trees, Depth, Criterion



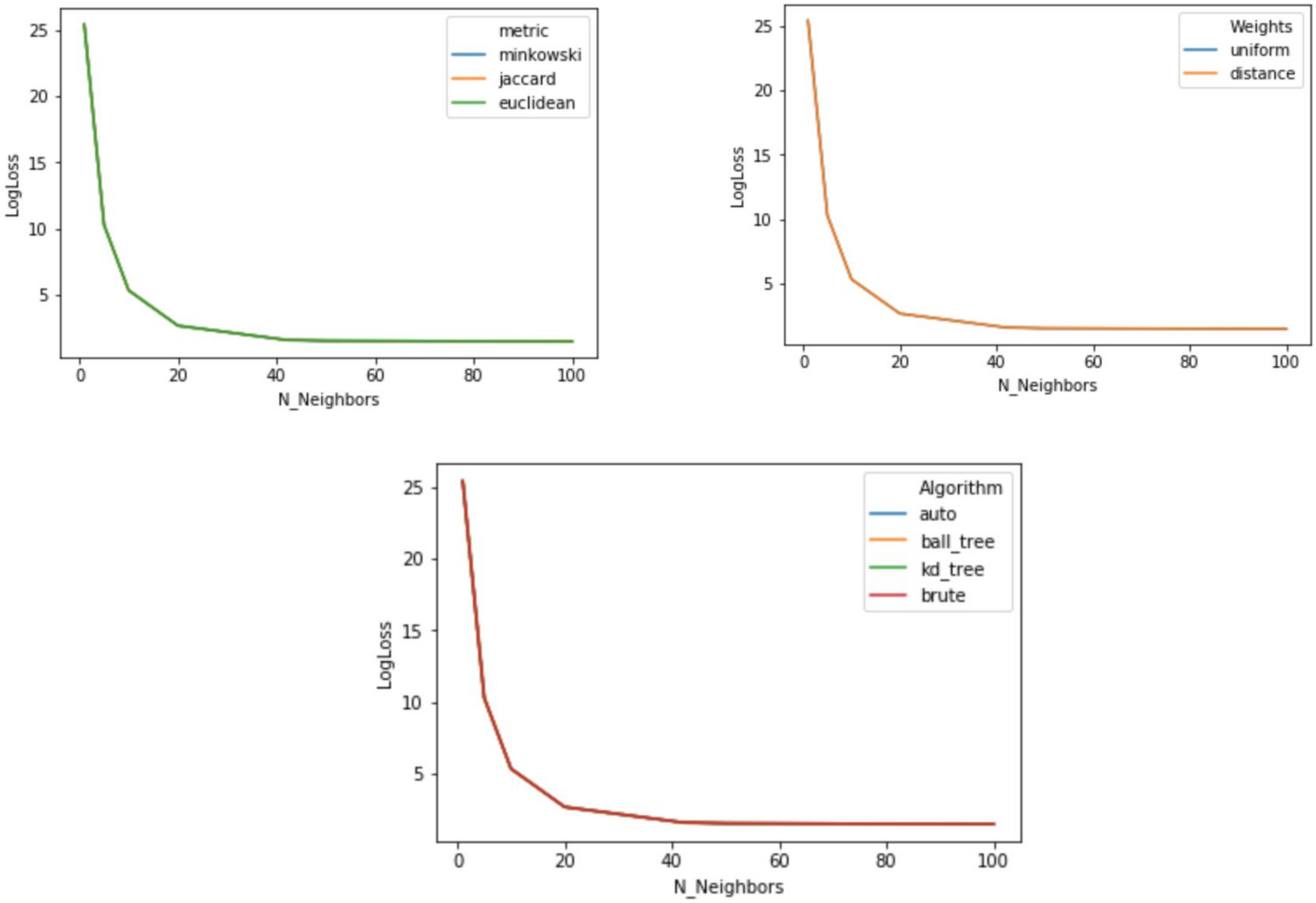
Appendix 2. Random Forest Tree Model for Depth:5, Trees=60 & Depth:3, Trees=60



Appendix 3. Average LogLoss for Gradient Boosting Tree Classifier - Trees, Depth, Learning Rate



Appendix 4. Average LogLoss for K_neighbors in K-Nearest Neighbors Classifier - Metric, Wights, Algorithm


metric
minkowski
jaccard
euclidean
LogLoss
N_Neighbors
Weights
uniform
distance
LogLoss
N_Neighbors
Algorithm
auto
ball_tree
kd_tree
brute
LogLoss
N_Neighbors