Using Machine Learning Techniques to Identify Digitally Excluded Individuals Across the City of Westminster, London

September 3, 2021

Abstract

Westminster City Council have set goals to increase the proportion of individuals accessing the internet across the borough. They hope to do this by setting up facilities where people can be introduced to, and taught how to fully utilise, important capabilities of the internet. In order to do this they first must understand which individuals they are targeting, and where these individuals are, so as to maximise the effectiveness of the program. This project has been completed with Westminster City council to help aid identification of individuals who may not have access to the internet. This is completed through the utilisation of data sets of survey responses and geo-demographic information provided by Westminster City Council.

Data was analysed using elastic-net regression and k-modes clustering with limited success. High error was found throughout due to high-dimensionality of feature vectors, missing data and limited numbers of samples. This report details in depth discussions of attempts to overcome issues relating to missing data, high-dimensionality and imbalanced data sets. Despite this, age, ethnicity, feelings of isolation, level of financial concern and job status were found to be significant indicators of limited or no internet access. An attempt was made to produce a simple score to indicate whether an individual is or is not digitally excluded using these indicators. Following tests of performance that output a mean recall value of 0.304 ± 0.111, this score was deemed unsuitable for use.

Further analysis of k-modes clustering and elastic net regression outputs suggests that individuals living in the north-west, south and east of the borough are most likely to be digitally excluded. Digitally excluded individuals are more likely to be older, facing social or economic hardship or of South Asian ethnicity. Despite attempts to mitigate their effects, all models fitted are susceptible to some level of use bias and data over-fitting.

Acknowledgements

anon

Contents

List of Figures

List of Tables

Nomenclature

AUROC Area under the reciever operator curve

COA Census output area

DE Digitally excluded

GAPS Glasgow Admissions Prediction Score

MICE Multivariate imputation by chained equations

NDE Not digitally excluded

ONS Office for National Statistics

PCA Principle Component Analysis

SMOTE Synthetic minority oversampling technique

WCC Westminster City Council

WPAG Westminster Population by Age and Gender

WPD Westminster Paycheck Directory

1 Introduction

Existing literature suggests that digital inequalities reflect, if not augment, any existing social inequalities in a population. A 2014 survey of the Dutch population demonstrated that individuals with higher education, higher paid jobs and younger individuals were receiving benefits such as better education opportunities, more and better health information and opportunities to acquire more friends or romantic relationships. From an economic perspective, those with higher capital, traditional members of the "upper class" can utilise advanced digital skills to trade online in stock markets [Piketty, 2018], or even gain professional relations online through specialised social networks such as LinkedIn [Helsper et al., 2015]. In general individuals from the upper and upper middle classes benefit from significant quantities of online contacts such as colleagues, acquaintances or customers whom may provide a valuable informational resource to greater increase socioeconomic well being [Poushter et al., 2018].

Looking to the future as artificial intelligence (AI) and internet of things (IOT) devices are increasingly prevalent and able to complete more lower-skilled tasks such as cleaning, construction and general manual labour, it is the lower socioeconomic classes that will increasingly struggle if they are unable or unwilling to adopt advanced digital skills [Mckinsey]. With this knowledge and motivation, Westminster City Council (WCC) are developing strategies to identify and surmount digital inequalities that exist across the borough. This research project is aimed at identifying individuals most at risk of digital exclusion in order to direct the council's resources to areas where they are most needed.

WCC have provided data sets including geo-demographic information about individuals across Westminster. These include a city wide survey about life in Westminster, Ofcom data about internet connectivity as well as various other data describing demographic and cultural information and material well being. It is hoped that, through the use and amalgamation of these information sources, a greater understanding of who is at risk of digital exclusion and where these people are will be gained. The approach taken in this masters project is to create a 5-15 item checklist of indicators of digital exclusion derived through logistic regression analysis of the data sets. Each item on the checklist will have a score associated with it. The end summation of these scores will provide a vulnerability score with a binary outcome of likely to be digitally included or not likely to be digitally included.

1.1 Project Aims

There are two main aims to this project:

- **To provide WCC with guidance on the characteristics of individuals that may be digitally excluded.** WCC have set goals to increase the proportion of individuals accessing the internet across the borough. They hope to do this by setting up facilities where people can be introduced to, and taught how

to fully utilise, important capabilities of the internet. In order to do this they first must understand which individuals they are targeting, and where these individuals are, so as to maximise the effectiveness of the program. Utilising raw data provided by WCC, analysis has been undertaken to fulfil this identification and thus this aim.

- **To provide WCC with ideas and information on how to approach the analysis of both qualitative and quantitative survey responses with machine learning.** The majority of the useful data provided by WCC is in the form of survey answers. Providing a pipeline of actions to be undertaken such that survey answers may be efficiently and accurately analysed using machine learning techniques will allow a greater understanding of any data collected as well as significant time saved.

All actions and discussions in this report both directly and indirectly work to achieve these two aims.

1.2 Report Structure

Section 2 provides an overview of current literature regarding digital exclusion, as well as a brief discussion on current vulnerability scores created in the medical sector. This is followed in section 3 a by a discussion of key machine learning theories and techniques that will be useful throughout the project and referenced in nearly all of the experimental sections. Sections 5 to 9 then discuss all research undertaken throughout the masters project. Each section discusses a small part of the experimentation and often takes the form of a self contained report, discussing theories and method relevant to the part of the experimentation described in the section's title. All experimentation has associated code. These may be found in the appendix (the source code uploaded via. Keats). Prior to an overall project conclusion, a brief discussion of legal, social, ethical and professional issues can be found in section 10. The research is then brought together in a discussion of potential future research and project-wide conclusion in section 11.

2 Literature Survey

This section discusses previous research undertaken, both in the field of digital exclusion and regarding vulnerability scores. The digital divide section leans heavily on the work of Jan van Dijk, arguably the most prevalent figure in the field of digital divide research. Unfortunately a majority of his work focuses on digital exclusion/inclusion in Holland, therefore it may not be directly transferable to a UK demographic. As such, an attempt has been made to include statistics relevant to the UK and Ireland wherever possible.

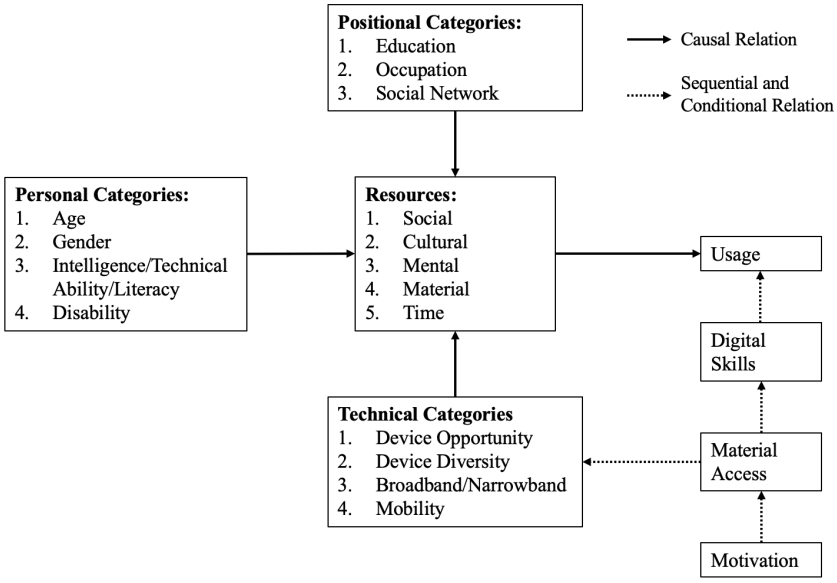


Figure 1: Factors affecting digital inclusion, adapted from [Van Dijk, 2020, Figure 6.2].

The vulnerability score section gives a brief overview of the use and creation of some vulnerability scores in existing literature. An in depth explanation of the mathematical theory involved is reserved for sections 7 and 8 and is introduced where relevant.

2.1 The Digital Divide

In a significant proportion of literature around the topic, digital inclusion is synonymous with the *digital divide*. In [Van Dijk, 2020], the digital divide is defined as

> A division between people who have access and use of digital media and those who have not.

The reasons for a digital divide are numerous with significant research being undertaken on the topic since the advent of computers. Fig. 1 lists many of the factors and causal relations effecting the digital divide and access to digital media. It provides a concise overview of many of the reasons for a digital divide. Whilst this report will focus on the causes of the digital divide rather than its effects, it is important to note why this research is necessary. A 2020 UK parliamentary report examining the effect of the digital divide during the COVID 19 pandemic highlighted that digital exclusion hinders an individual's ability to readily access health and social care information as well as online employment support and job applications. Furthermore, well-being

activities such as online gym classes or contacting family and friends were almost impossible without internet access throughout much of the pandemic [Baker et al., 2020]. Post-pandemic, the country is expected to return to a "new normal" with an increase in online activities [Harper, 2021]. This will further augment any existing digital divide if the problem is not sufficiently tackled.

The focus of digital divide research has changed significantly since the 1990s. In Western European and North American countries, focus has shifted from physical access to usage and skills. The following section addresses both of these topics, including a discussion of which appear to be more or less relevant today.

- **Personal Factors**: In the early days of the internet there were clear divides in age and gender of people using and up-taking digital media, with young men being the quickest to gain skills with the elderly many years behind [Thayer and Ray, 2006]. Today there is still a level of age division with a 2020 office for national statistics (ONS) survey stating 100% of respondents under the age of 35 accessed internet daily or almost daily, this figure fell to only 67% in the over 65s [ONS]. Further research conducted in Sweden suggested that the range of internet usage in older individuals differed greatly between the "younger old" and the more elderly (80+) [Berner et al., 2015]. Further ONS data suggests 58% of internet non-users and 61% of those claiming to have zero digital skills in 2018 were women. Some studies have shown, however, that men are more likely to overstate their digital competence when compared to women thus it is important to incorporate empirical and controlled sociological data into survey response analysis [Van Dijk, 2020].

  Possibly one of the least researched, significant factors in internet access is disability. A 2016 survey of Americans concluded 8% of individuals have never accessed the internet, this figure rose to 23% among disabled individuals. When enforcing controls to account for age, gender, employment and education level, this effect of digital exclusion of disabled people remained [Van Dijk, 2020]. Velleman 2018 demonstrated that many interfacing aids and web adaptations for those with further needs are under developed with organisational guidelines often not adhered to.

  Literacy is also a barrier to digital inclusion. An estimated 8 million adults in the UK are functionally illiterate [Ullah, 2015]. Whilst the rise in primarily visual media such as photos or videos has allowed for some to successfully access and navigate the internet, one may still be digitally excluded from more involved tasks such as online healthcare or job applications.

- **Positional Factors**: Perhaps correlated with literacy and ability is education. Consistently, those with higher levels of education have higher and differing use levels on the internet to those who have lower levels of education. From a digital skills perspective, those with higher education levels, or jobs that require the

use of computers, perform consistently perform better in laboratory assessments of content related skills and decision making [van Deursen and van Dijk, 2015].

Further to occupation and education, the company one keeps can influence internet access and behaviours. For older adults, if they live alone they are significantly less likely to obtain internet access than their counterparts residing in multiple person households [Doody et al., 2020]. Outside of the household, it has been recorded that those with larger social networks are the first to receive strategic information from their acquaintances, this in turn leads to a mirroring between the social and digital elite, potentially diminishing prospects of social mobility [Kadushin, 2012].

- **Technical Factors**: According to Ofcom, 98% of UK households have the physical infrastructure for a 10MBits^{-1} download speed and 1MBits^{-1} upload speed internet connection [Ofcom, 2020a]. This was deemed a "decent" internet speed. Despite near universal availability of basic internet access and hardware, the digital divide still remains. An individual with up to date hardware and fast internet will benefit more from it than someone with older, more unreliable technology. An individual with sole access to a smartphone or tablet may not have the mobility of skill to use more involved applications on a desktop computer. Furthermore, sole smart phone use limits content and memory access for many applications thus is generally insufficient for an individual's complete digital inclusion [Napoli and Obar, 2014].

Due to the Ofcom asserted availability of 'decent' internet speed in the majority of households, it is expected that absolute access to infrastructure required for internet access will not be an issue across Westminster. Instead focus shifts towards the demographics of people accessing, or not accessing, the internet. A majority of previous research has been undertaken outside the UK and as such cultural differences, however minimal, may lead to disparities between current data and previous research. In addition attitudes towards the internet may have changed significantly since the Coronavirus pandemic. Data acquired for this MSc research was collected in late 2020, almost a year since the Coronavirus act 2020 [UKGovernment, 2020] introduced emergency restrictions on individuals to prevent the spread of Coronavirus.

2.2 Vulnerability Scores

Vulnerability scores are utilised across many disciplines as simple indicators of binary outcomes. A vulnerability score is a score defined by a series of data about an individual in order to predict the classification of an individual as vulnerable or not vulnerable to a given outcome. For example, The Glasgow Admission Prediction Score (GAPS) [Cameron et al., 2015], is a score created to identify the likelihood of hospital admission when an individual enters an accident and emergency department in hospital. The score accounts for factors such as gender, age and previous health

conditions. The influential factors on admission were identified using a multivariate logistic regression model trained on 215231 samples. Following formulation and validation on 107615 samples, GAPS successfully outperformed nurses when making specific binary predictions [Cameron et al., 2017]. Similar predictive scores are used across medicine with multiple examples available across many different fields including Liang et al. 2020, Kim et al. 2016, Leo et al. 2019. Despite differences in application, the development of many of these scores follows a similar process utilising logistic regression for primary analysis.

An attempt will be made to transfer the process of vulnerability score creation across to the Westminster City Council data. Unfortunately sample sizes of council data are often little over 1000 and therefore adaptations must be made in the creation of vulnerability scores to account for the minimal data. The creation of vulnerability scores using minimal data has not been found in this literature survey. It is hoped that under and oversampling techniques, as discussed in section 6 will overcome this issue and provide useful insight into the creation of well-performing vulnerability scores with limited data. Furthermore, the adaptation of the vulnerability score technique away from the medical field provide insight into the flexibility of such methods.

3 Machine Learning Theories

The following section discusses machine learning theories important for use throughout the entirety of this report. Other theories and background information about algorithms will be introduced in specific sections when relevant.

3.1 Logistic Regression and Regularisation

Logistic regression is a technique used for both binary classification and regression. It extends techniques developed in linear regression through the incorporation of probability metrics defining the likelihood of an individual sample being in a given class. A sigmoid function represents probability of data relating to one of two classes and is fitted to the data. If the probability is above a threshold value then it is classed as one class, if the probability is below this threshold value, it is classed as the other.

The traditional form of a logistic regression model takes the probability of point i being categorised in a given class as

$$p_i = \frac{\exp\{(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in})\}}{1 + \exp\{(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in})\}} \tag{1}$$

where β_0 is the bias of the model, β_n is the coefficient of the n^{th} feature x_n. This can be modelled as a linear response using

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in} \tag{2}$$

where ln is the natural logarithm.

When creating a logistic regression model, the objective is to create a model that maximises the likelihood of the model predicting true outcomes. This translates to an attempt to maximise the log-likelihood function, equal to

$$\ln(L) = \sum_{i=1}^{N} \left[\ln(1 - p_i) + y_i \ln \left(\frac{p_i}{1 - p_i} \right) \right] \tag{3}$$

where y_i is the true classification of point i and N is the total number of points in the data set. In reality computational efficiency is increased with the goal of minimising a log-loss function,L_{log}, equal to the negative of the log-likelihood. The maximum likelihood is then found through the minimisation of the negative summation of relevant probabilities using gradient descent [Hosmer Jr et al., 2013]. This works well for some data, however, can have performance significantly hindered by overfitting. In addition, if the dimensionality of the feature space is high, extremely complex and convoluted models may be created. Both these problems may be overcome through the introductions of L1 and L2 penalties to the model in the form of elastic-net regression.

The first penalty introduced is the L2 penalty, creating a log-loss function of the form:

$$L_{log} + \lambda \sum_{j=1}^{n} \beta_j^2 \tag{4}$$

where j is the number of features in the data. The penalty introduced is β_j^2 and λ is a hyper-parameter to define the weight of the penalty. The L2 penalty introduces a small amount of bias in order shift the fitting of the model away from too closely mapping to the training data and thus over-fitting. Every model parameter except the original logistic regression bias term β_0 is included in the L2 penalty, thus changing solely the gradient of the L_{log} function for minimisation. Furthermore the L2 penalty the issue of using a high dimensional feature space with limited data points by making predictions less sensitive to the training data.

The second penalty introduced is the L1 penalty. This penalty takes the form:

$$L_{log} + \lambda \sum_{j=1}^{n} |\beta_j| \tag{5}$$

[Podgórski, 2021]. The forms of both of these penalties are very similar and can be used in very similar situations however they have slightly different capabilities. The largest difference between the two capabilities is that the L1 penalty has the capacity to shrink the contribution of a feature to zero whereas the L2 penalty can only shrink the contribution asymptotically close to zero. This is very useful with data such as the WCC survey data used in this report, when it is not clear how useful individual features may be when predicting digital exclusion. The L1 penalty increases computational efficiency when creating a model. Despite the benefits of the

L1 penalty, to individually discard parameters completely may be premature. Thus combining the L1 and L2 penalties is the best option.

The penalties are combined in elastic net regression to introduce a log-loss of the form:

$$L_{log} + \lambda \sum_{j=1}^{p} (\alpha\beta^2 + (1-\alpha)|\beta_j|) \tag{6}$$

where *alpha* dictates the ratio between the L1 and L2 penalties.

Elastic net regression performs well when there are correlations between features. L1 regression alone may pick a single feature out of a correlated set to discard and ignore others, leading to little increase in performance. Elastic net regression groups and shrinks or removes contributions of correlated feature and is thus chosen as the preferred regression technique for the WCC Survey data. [Zou and Hastie, 2005]

The introduction of regularisation introduces many hyper-parameters to be evaluated in order to find the best performing model. The evaluation of hyper-parameters is undertaken using k-fold cross-validation. k-fold cross-validation splits data into k equally sized subsets. k models are built using k-1 subsets for training the model and a single subset for testing the model. The subset used for testing the model is changed for each model built. Once all k models are built, their performances are evaluated. The next iteration of model building changes a single hyper-parameter in order for a direct comparison of the effects of changing the given hyper-parameter. Once this process is undertaken for all possible hyper-parameter combinations, the best model is found and used for the final result [Fushiki, 2011].

3.2 Performance Metrics

There are numerous metrics that can be used to test the performance of a machine learning model. Each of these metrics tests a different element of performance and should be used for different applications. Throughout this project classification models are built to identify whether an individual, represented by a data sample, is, or is not, digitally excluded. Digital exclusion (DE) is represented as the positive class throughout. Most machine learning metrics are based upon four measures recorded after classification. These are:

- **True Positives** TP: The number of samples correctly identified as belonging to the positive class.
- **True Negatives** TN: The number of samples correctly identified as belonging to the negative class.
- **False Positives** FP: The number of samples belonging to the negative class but classified as belonging to the positive class.
- **False Negatives** FN: The number of samples belonging to the negative class but classified as belonging to the negative class.

Across all data sets provided by WCC, a severe imbalance exists between the number individuals who are digitally excluded and those who are not. Across all data sets, when target data was created, the average rate of digital exclusion was $11 \pm 4\%$. The equation for accuracy is

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

. When using unbalanced data, a classifier that classifies all samples as belonging to the negative class could have deceptively high accuracy without providing any useful information [Jeni et al., 2013]. As such accuracy is deemed inappropriate for use as a performance metric.

Precision and recall are two metrics based on the proportion of classifications of positive class that are correctly identified. Precision takes the form

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

and describes what proportion of samples classified as positive are classified correctly. Recall takes the form

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

and describes the proportion of actual positives that are classified correctly [Buckland and Gey, 1994].

In this report it is important to correctly identify all individuals who are digitally excluded. If an individual is incorrectly identified as digitally excluded then they can simply decline to access the resources WCC may provide. However, if an individual incorrectly identified as not digitally excluded, then the individual may not be able to access useful resources leading to an augmentation social or economic repercussions of not being online. Therefore, recall was chosen as a primary performance metric to maximise in this report due to its minimisation of false negatives.

Despite recall being utilised as the metric of most importance in this report, precision should not be totally ignored. If too many individuals are falsely identified as digitally excluded this may lead to a large over-estimation of the number of resources needed. This overestimation may lead to both financial, temporal and physical waste. F1-score is the harmonic mean of precision and recall and is useful as a metric when attempting to balance the two. F1-score is of the form:

$$F1Score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{10}$$

A third measure of performance is introduced alongside recall and f1-score. The area under the receiver operator curves (AUROC) evaluate performance when using different thresholds of the logistic regression classifier. The logistic regression classification threshold is a value between 0 and 1, all samples above the threshold are classed as

positive and all sample below the threshold are classed as negative. A graph is created with the true positive rate $\frac{TP}{TP+FN}$ on the y axis vs false positive rate, $\frac{FP}{FP+TN}$ on the x axis for each threshold. The AUROC is a value between 0 and 1, 1 being a perfect model. The AUROC is introduced as a performance metric as it is independent of the classification threshold. This allows for a true idea of the quality of a model irrespective of where the threshold is placed. [Hanley and McNeil, 1982].

4 Available Data

Multiple data sets are provided by WCC. These include both survey answers and general demographic information by postcode. A description of each data set and its inclusion in this study is provided in this section.

4.1 Westminster City Council Survey

The majority of this project will focus on the WCC Survey for 2020. The WCC Survey asks 45 questions about many facets of life in the borough. Questions are mainly directed to identify potential areas for improvement for the council such as refuse collection or street safety. Answers to these questions provide feature and target data for input into the vulnerability score. 1038 sets of survey answers are collated in an excel file in the form of both numerical and categorical data. Further information on how this data is used can be found in section 5.

4.2 Quantitative socioeconomic data by postcode

30794 individual postcodes exist across Westminster with data provided about internet connectivity, income and age and gender of residents for each postcode. Of these postcodes, 5138 have a total residential population of 0 so have been discarded. The highest total population of any postcode is 404, with a significant proportion of postcode areas having a population of 1. Below is a list of data sets providing socioeconomic data by postcodes. These data sets were combined for input into a location-based clustering algorithm as discussed in section 8.

- **Westminster Population by Age and Gender Directory** (WPAG): This data set lists a total of 30794 postcodes across Westminster of which 17545 have been immediately discarded as large users (a company building which receives 500+ items of post daily). This data set divides individuals by gender and age brackets. Age brackets are on average five years in length. An integer value states the number of individuals of a given gender in each age bracket.

- **Westminster Paycheck Directory** (WPD): This data set states absolute income for individuals at a given postcode as well as mean, median, mode and lower quartile income.

Category	*Group*		*Type*
1. Affluent Achievers	A	Lavish Lifestyles	1-3
	B	Executive Wealth	4-9
	C	Mature Money	10-13
2. Rising Prosperity	D	City Sophisticates	14-17
	E	Career Climbers	18-20
3. Comfortable Communities	F	Countryside Communities	21-23
	G	Successful Suburbs	24-26
	H	Steady Neighbourhoods	27-29
	I	Comfortable Seniors	30-31
	J	Starting Out	32-33
4. Financially Stretched	K	Student Life	34-36
	L	Modest Means	37-40
	M	Striving Families	41-44
	N	Poorer Parents	45-48
5. Urban Adversity	O	Young Hardship	49-51
	P	Struggling Estates	52-56
	Q	Difficult Circumstances	57-59

Figure 2: Table describing all Acorn category types as well as their individual type codes [CACI, 2019b, pg.3]

- **Ofcom Connectivity Data** This open-source data set provides details of the internet connectivity speed across postcodes in Westminster. For each postcode, the percentages of households with download speeds of less than 30Mbitss^{-1} and those with 1Gbitss^{-1} download capabilities are provided [Ofcom, 2020b].

4.3 Acorn Directory

The Acorn directory divides the UK population by postcode. Each postcode is assigned a category label based on demographic data, social factors and population and consumer behaviour combining multiple data sources including, but not limited to, open-source, government and commercial data. Category assignation is undertaken using machine learning then enhanced using human research panels. Using Acorn data and attributes associated with each designated category, a good idea of the attitudes and prosperity of individuals residing in a given postcode will be obtained [CACI, 2019a]. Each postcode is assigned a single integer number between 1 and 59 as shown in fig 2. These integer assignations are used in both the vulnerability score and the location-based clustering algorithm in order to greater understand attributes of individual data samples.

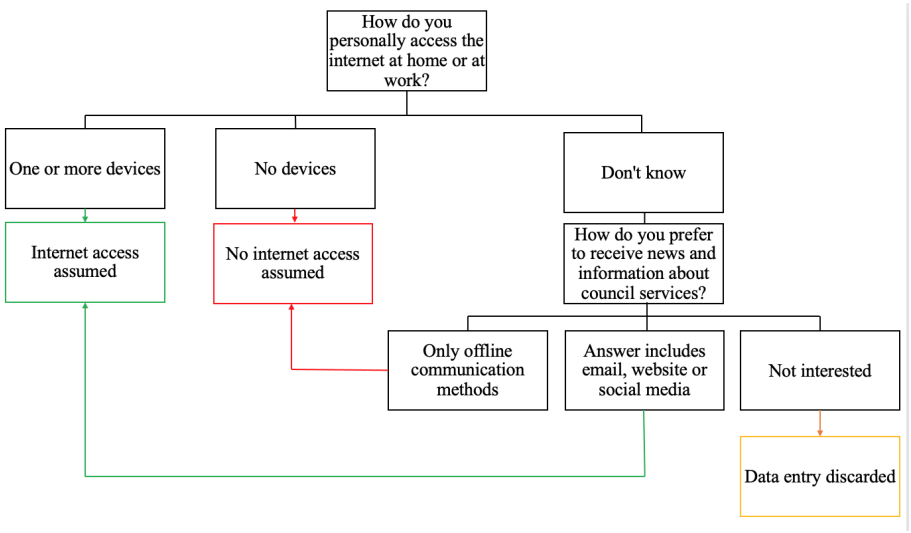


Figure 3: A flow chart of how target categories are decided for Westminster City Council Survey Data

5 Data Cleaning and Feature Engineering: Vulnerability Score Input

The primary focus of this report is the WCC survey for 2020. Unfortunately upon initial examination of the data, the data was seen to have significant quantities of missing values, as well as multiple questions irrelevant to digital exclusion. This section discusses the data cleaning and engineering methods implemented such that the survey data is both relevant and suitable for input in a machine learning algorithm.

5.1 Target Creation

There are limited questions directly referring to internet usage and a combination of these has been used to split data into two categories, *not digitally excluded*, NDE, and *digitally excluded*, DE. Fig. 3 demonstrates how target categories are decided. When internet access is assumed, an individually is classed as NDE. Otherwise they are classed as DE. With reference to fig. 3 if an individual does not know on what devices they access the internet, further differentiation is required through the investigation of how they wish to receive news. This design decision is highly subjective and premature discarding of data could have implications in the performance of any model built. However, once target data was decided using the process in fig. 3, it was found that no data entries were discarded. Using this definition of digital exclusion, 7.3% of 1038 individuals were found to be digitally excluded.

In addition to an assumption of total digital exclusion, further adaptation of the DE class of target data was undertaken to distinguish between those only accessing

internet via their smartphone and those using other devices. Smart phones provide a cheap and mobile method of internet access however often lack the memory capabilities and screen size to easily complete the more information intensive tasks often involved in education or working environments. Internet use on smart phones has therefore been shown to be mainly limited to entertainment and leisure with some research papers have highlighting the effect as creating a "mobile underclass" [Napoli and Obar, 2014]. All individuals classed as DE using dichotomizer in fig. 3 were discarded in order that they did not skew mobile exclusion target data. If an individual only accessed the internet through a mobile phone then they were categorised DE for this purpose. Using this definition, it was found that 14.35% of individuals who completed the survey were digitally excluded. Data that has a target formed in this manner will henceforth be referred to as mobile exclusion data.

5.2 Evaluating Data Importance

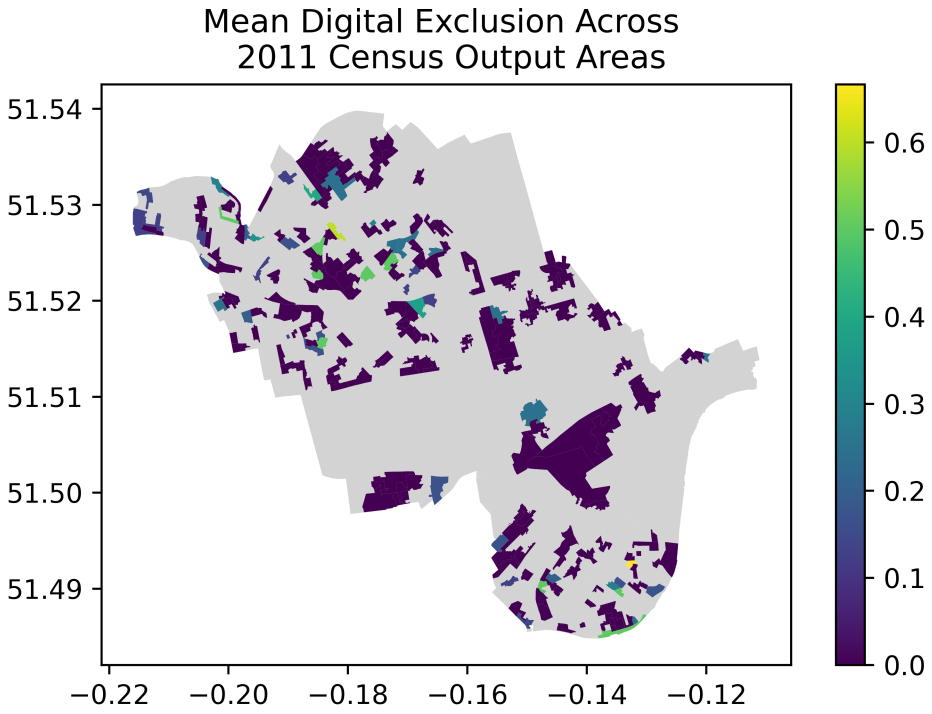
Of the 45 questions in the WCC survey, 20 were immediately manually discarded prior to input into a learning model. These included questions regarding how satisfied individuals were with the WCC's approach to refuse collection, pavement maintenance and more. These questions were deemed irrelevant in the context and could safely be assumed that any correlations found would be purely incidental.

After manual discarding of answers deemed irrelevant to the model, the feature space of the WCC Survey data still possessed extremely high dimensionality. In order to decrease processing time of the model, this feature space was further reduced through combining answers to logically correlated questions or combining answers to multi-answer questions into a single combined feature. For example, the question of *which of the following council services have you accessed?* was compressed into a single variable stating the number of council services accessed. This was deemed appropriate as all the listed options for council services involved services for those suffering from socioeconomic hardships.

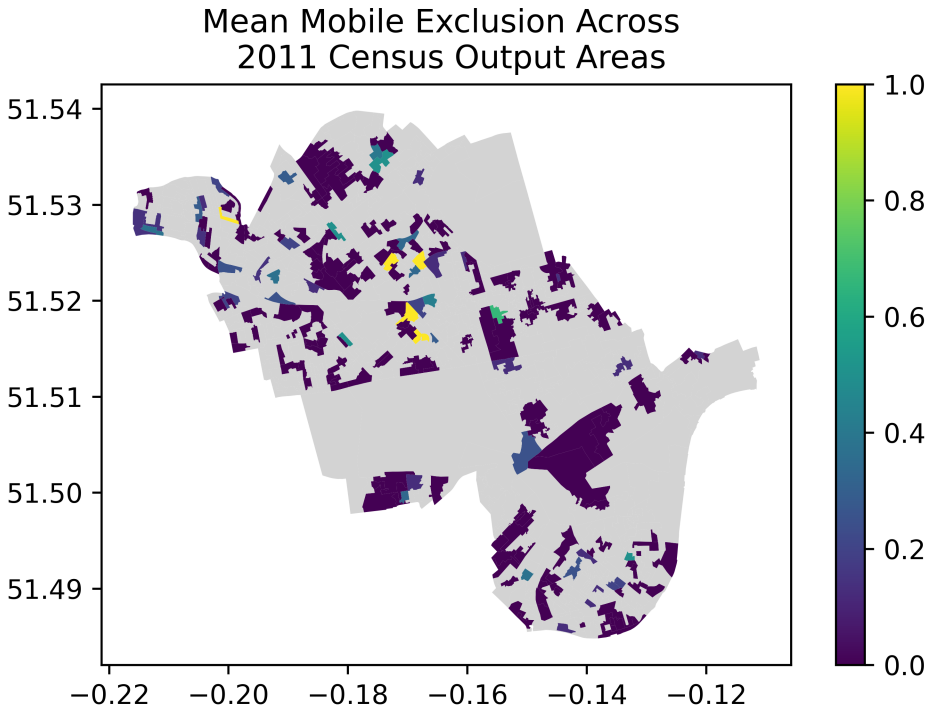
5.3 Categorical Data

The nature of the WCC Survey led to significant quantities of features containing categorical data. Some machine learning algorithms have capabilities for handling categorical data however the majority implement some kind of distance estimate for training which is reliant on numeric data. In addition, categorical data means nothing to an algorithm and as such will often not greatly improve performance. Further investigations into the use of natural language processing for the prepossessing of survey responses could prove useful for future survey analysis, however, due to limited project time, it was deemed more efficient to manually explore and redefine categorical data.

The categorical data with the maximum number of possible values was postcode



(a) Plot of total digital exclusion



(b) Plot of mobile digital exclusion

Figure 4: Maps of the borough of Westminster showing the proportion of individuals facing digital exclusion in each census output area. Grey colours symbolise no data available

data. This was encoded numerically using the *pgeocode* package [Yurchak, 2020]. The mean latitude and longitude values were taken for each postcode and input into the features space. Fig. 4 shows mean digital exclusion per census output area (COA) across Westminster. COAs are areas incorporating around 125 households with a population of around 300. COAs were deemed a good metric upon which to plot location data due to their relatively small scale, whilst maintaining a level of simplicity and larger scale visualisation not seen when using individual postcodes [ONS]. As seen in fig 4, the location of WCC Survey respondents is sporadic for both total and mobile digital exclusion data. The inclusion of location data when building models with the WCC survey data hindered model convergence. Due to the level of incompleteness of the location data, it was deemed inappropriate to attempt to impute the data in order to get a full idea of the influence of location when creating the vulnerability score. As such quantitative socioeconomic data by postcode (section 9) was analysed to give some indication of the location of digital exclusion across the borough.

If the answer to a question had two possible answers, the two options were quantified with a Boolean truth value of 0 or 1. Multiple choice answers dealt with in two possible ways. If there was some quantifiable pattern across the answers then each answer was assigned an integer value according to the pattern. For example if a question asking about how often an individual felt lonely had the answers, *not at all, some of the time, most of the time*, then each of the survey responses would be assigned values of 1,2 and 3 respectively. Following integer assignment of values a dictionary was created to link integers to their original values. This dictionary was saved as a .json file for later reference.

If there was no quantifiable pattern across multiple choice answers, for example, options for ethnicity, the feature was encoded using one-hot encoding. One-hot encoding assigns Boolean truth values to each possible value of a feature. If a feature has n possible answers then n new features will be created, headed with the feature value. The values of the new feature are posed as a Boolean truth array [Yu et al., 2020].

5.4 Data normalisation

Machine learning models may become skewed and wrongly weight features of higher or lower importance if feature values are of significantly different magnitudes. As such, prior to input into any machine learning model, all feature values were normalised. This was undertaken by finding the maximum value for a feature and dividing all other values for the given feature by this maximum value. In addition to normalisation, this technique identified any features where there was a single answer given across the feature array. As these features would not aid model training, they were immediately discarded.

5.5 Missing data

All answers to questions stating "Don't Know" or "Prefer not to Say" were classed as none types in the data. All the none-type data items were then counted and if a feature had a greater than 25% occurrence of none-type values it was investigated to see if the data shortage may be experimentally significant. Significance testing was undertaken by taking the mean of the values of the target variable in each row with a none-type data item and comparing it to the mean of the values of the target variable in each row with a none-type data item. Following this analysis features stating peoples feelings about losing their business, returning to school, and the perceived usefulness of selected online publications were discarded.

Despite removing features with a significant proportion of missing values, 25 features remained with a missing values. Of these features, 9 had less than ten missing values and four had over 100 missing values. These missing values prevent many model building algorithms, including logistic regression, from building good models. Thus, an attempt was made to impute the missing values using multivariate imputation by chained equations (MICE).

5.5.1 Iterative Imputation of Missing Values

It is expected that much of the missing data in this survey is due to the respondent not wishing to disclose information. If a respondent does not respond to a survey question for a given reason, for example, not wanting to disclose financial hardship, it is likely that others may similarly not respond. This will skew any imputation that may be undertaken and as such imputed data will be less reliable. Despite this, due to the quantity of missing values in the data, imputed data may yield better results than simply discarding survey responses with missing fields. The following section discusses an investigation into the effectiveness of different types of iterative imputation when compared with none-imputed data.

MICE takes the feature data and in turn sets each column with a missing value or values as the target variable and builds a regression model to predict the target. This is repeated multiple times and the results combined until convergence. Prior to the initial model build, data is imputed by replacing missing values with the mean value of the feature. A model is then build in turn for each feature with a missing value or values. The feature with lowest quantity of missing data is imputed first.

Following each iteration, i, of imputation using machine learning models, the difference between the i^{th} value in the feature data and the $(i-1)^th$ value is calculated. If the difference is below a threshold value, the value has converged and imputation is halted, otherwise a new model is created from the i^{th} data set. This process is repeated until all values in the feature set have converged or a predefined number of iterations have been undertaken [Buck, 1960, Van Buuren and Groothuis-Oudshoorn, 2011].

There are many possibilities for the algorithm used for imputation. These include:

- **Bayesian Regression**: Bayesian regression builds on traditional linear regression (fitting a best fit line or plane to data points). Instead of utilising individual data points to fit the model, target data is is generated from a Gaussian Distribution based on all samples. The mean of the distribution is the transpose of the matrix of coefficients output by a simple machine learning model. The variance is the square of the standard deviation of this mean. [Koehrsen, 2018]

- **Decision Tree Regression**: This method splits each feature into two or more branches dependant on the value of the feature. A base (root) point (node) of the decision tree is formed from the most divisive feature. The values of feature form the leaf nodes of the tree. From each leaf node, a new feature node forms specifying the next feature. This feature node sprouts leaf nodes with its values. This process of node creation continues until all features are utilised or the target can accurately be predicted from paths through the tree [Clémençon and Vogel, 2020].

- **Extra Trees Regression**: This method is an ensemble method utilising many decision trees. Extra trees utilises all the data to build a tree then randomly splits nodes at cut points to create multiple decision trees. The final value for the target is decided by taken the arithmetic mean of the output of the trees.

- **K Neighbours Regression**: This method finds the K nearest neighbours of a sample point in order to identify the sample class. K is an integer assigned by the user. Nearest neighbours are identified using a Euclidean distance metric for all numeric variables. The Euclidean distance metric is of the form

$$distance = \sum_{i=1}^{k} \sqrt{x_i - y_i} \tag{11}$$

  , where x and y are different samples in the data set and k is the total number of samples. Samples with the smallest distance between each other are classified as nearest neighbours in a class. The centre point of this class is taken to be a good estimate for imputed values [James et al., 2013].

An *impute* function was created to impute data and allow for comparisons between models created using data imputed with each of the different regression algorithms. The imputed data was then input into a Bayesian Regression algorithm. 5-fold cross validation was used for comparison of models and mean squared error was used as a performance metric. The imputation method with the lowest mean squared error was taken to be the final technique used to impute the data for the rest of experimentation.

The results of cross validation are shown in fig 5. There was a 0.018 ± 0.0077 reduction in mean squared error when comparing imputed and not imputed data. However the regression algorithm used was deemed inconsequential on model performance. Across the four regression techniques tested, there was a standard deviation on mean

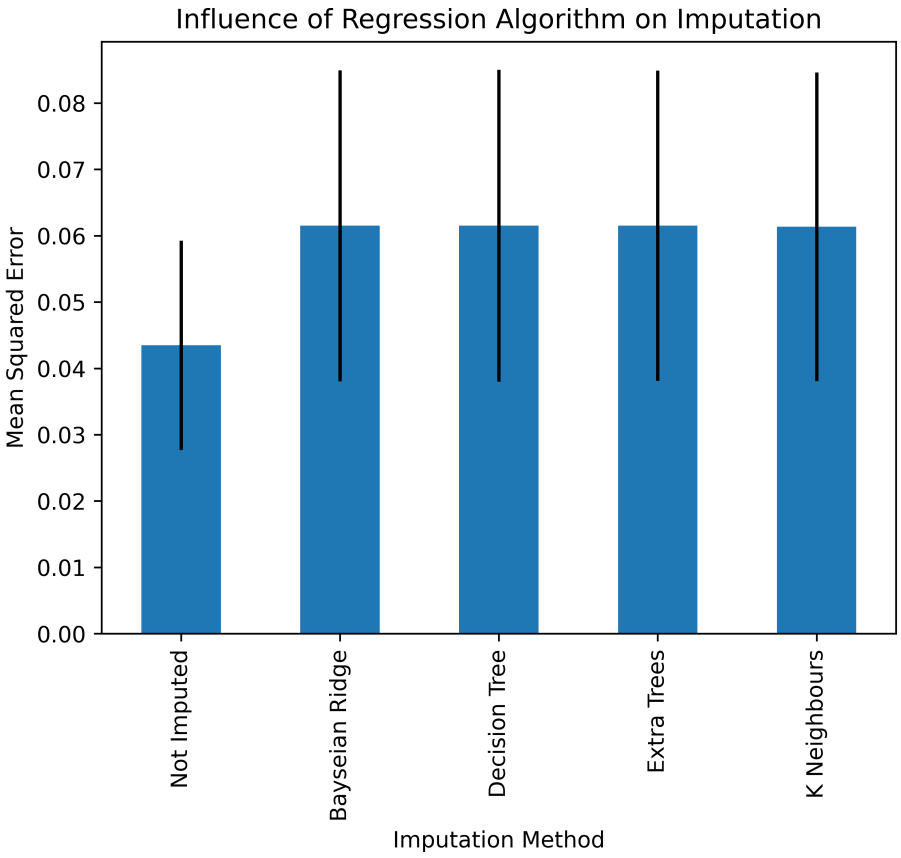


Figure 5: Bar chart to compare use of different regression algorithms when imputing data.

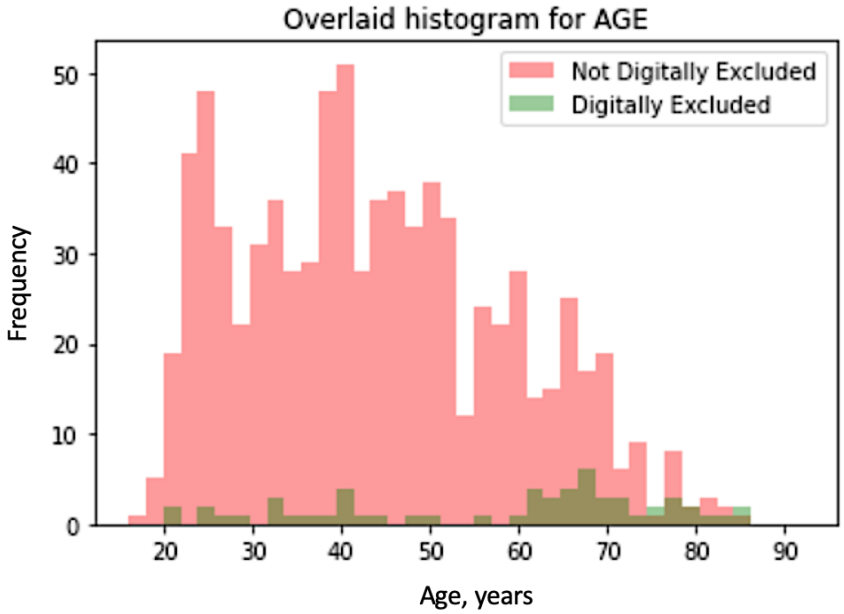


Figure 6: Histogram showing the spread of ages of individuals who are/are not digitally excluded using Westminster City Council survey data

squared error of 0.008. As the exact layout of the data is not known, it was decided that using the linear method of Bayesian Ridge regression may not accurately reflect the input data. K Neighbours regression can be heavily influenced by the random situation of cluster centres and was thus deemed risky to use as an imputation method. It has been shown that Extra Trees reduces variance and overfitting when compared to Decision Tree regression, thus this was the method chosen to impute data for the rest of the experimentation.

5.6 Preliminary Data Analysis

Some preliminary analysis of the data was undertaken to allow for early identification of influential features. This would allow for a more guided approach to later analysis. Preliminary analysis was undertaken using Jupyter notebooks. All preliminary analysis was undertaken after identification of the target variable but prior to conversion of categorical features or removal of any features. Continuous numeric features were plotted using overlaid histograms of DE vs NDE individuals. Most features followed a similar distribution for both classes, however the data for age showed a skew towards older ages in DE individuals. This is shown in fig. 6. Overall the modal population seems to be between the ages of 20 and 50, however, for the fraction of the population classes as DE, there is a peak in the absolute number of individuals between the ages of 60 and 80.

6 Data Sampling and Model Combination

Machine learning algorithms perform best to create representative models when there is a near equal quantity of data available for each target class. The nature of digital exclusion in 2021 is such that, across the three data sets used, there was an average rate of digital exclusion of $11 \pm 4\%$. This highly imbalanced data led to a lack of convergence when building models. For logistic regression models built using 10-fold cross-validation for hyper-parameter tuning models that did converge there was a minimum mean squared error of $34.98 \pm 2.82\%$ across the data sets. Thus further feature engineering was deemed necessary to modify the data sets and improve performance.

Various undersampling and oversampling techniques were contrasted and tested to find the most suitable algorithm. The undersampled data was then combined with the undersampling techniques and performance was reviewed.

6.1 Theory

Oversampling and undersampling are techniques used in turn to artificially create data or selectively remove data. Many techniques exist for both sampling variations, however, only the most common methods are utilised in this report. These methods are discussed below.

6.1.1 Undersampling

Undersampling is the process of selecting points from the majority class to remove from the model input. The undersampling methods investigated in this report are as follows:

- **Tomek Links** focuses on removing overlapping data points. First the distance between two points, a and b, belonging to different classes, is defined using a Euclidean distance estimator. The pair of points are defined as a Tomek Link if there is no point c such that:

$$\delta(a, c) < \delta(a, b) \quad \text{or} \quad \delta(b, c) < \delta(a, c) \tag{12}$$

  Any points that fulfil the criteria in 12 are deemed to be boundary points or noise and are thus removed [Ivan, 1976].

- **Neighbourhood Cleaning Rule** removes data points according to the value of its three nearest neighbours. Nearest neighbours are defined using a Euclidean distance estimate. For each point in a data set, if a point is in the majority class and two thirds of its nearest neighbors are in the minority class, the point is removed. Conversely, if a point is in the minority class and two thirds of its nearest neighbors are in the majority class, the majority class instances of the point's neighbours are removed [Laurikkala, 2001].

- **Near Miss** keeps the samples for the majority class where the average distance to the k closest samples of the minority class is the smallest. Distance is measured using a Euclidean distance measure. k is an integer value. Both k and the number of points left in the majority class are user defined [Lemaître et al., 2017].

- **Cluster Centroids** performs k-means clustering on the data belonging to the majority class. The algorithm is instantiated with random cluster centres. A new majority sample data set is created from the centroids of each cluster following clustering.

6.1.2 Oversampling

Oversampling is the process of selecting points from the minority class to duplicate or use as a base for the creation of new dummy points. The most common technique for oversampling is the synthetic minority oversampling technique (SMOTE). Initially a point is chosen at random from the minority data. Using Euclidean distance metrics, 5 of the point's nearest neighbours are found. A point along the connecting lines between the point and its neighbours is chosen at random as the position of the synthetic sample [Chawla et al., 2002].

6.2 Method

Different combinations of data selection techniques were used to find the most appropriate strategy for data engineering across the WCC survey data. Metrics discussed in section 3.2 were used. If metrics conflicted, for example, if a change in technique resulted in an increase in one performance metric and a decrease in the another, recall was used as a primary metric.

A sampling class was created using Python 3.8 to sample and fit data. Functions were created for each sampling method utilising the Imblearn package for Python [Lemaître et al., 2017]. Throughout the experiment logistic regression was used to create models. 5-fold cross validation was implemented to tune model hyper-parameters. The models created were based on training data sampled according to specified sampling methods. The model was then fitted to the original survey data, both training and testing sets. Performance was evaluated on both sets. The data used was the numeric data prepared as discussed in section 5. Principle component analysis (PCA) was used for data visualisation to display the spread of the two principle components of the WCC Survey data. PCA takes the two features with the most variance across the data set. These features are then plotted with axes representing the projection of the individual feature's maximum variance [Jolliffe and Cadima, 2016].

The first models created were created using the under sampling techniques discussed in section 6.1.1. Alongside the performance metrics stated above, the output ratio

before/after sampling was also considered. The output ratio is defined as the number of samples in the minority class divided by the number of samples in the majority class.

SMOTE was then implemented with varying output ratios, N, between 0.1 and 0.9. The performance of the models built under the different data conditions was recorded an evaluated.

Following a completion of under and oversampling, the best performing techniques from both methods were identified and combined in an attempt to further increase model performance.

Throughout the experiment, each sampling technique was performed three different times. The data input to each iteration of the sampling algorithm was equal to 70% of the WCC survey data split randomly using the Sklearn's inbuilt *train_test_split* method [Pedregosa et al., 2011]. A model was created for each iteration of the data sampling and it's performance evaluated. Final results were found based on the mean of the performance of these models with errors equal to the standard deviation of the mean.

6.3 Results and Discussion

Fig. 7 shows a comparison between the performance of a model built using data sampled using the named technique. For the Near Miss and Cluster Centroids methods, a ratio of 0.5 was used as an initial ratio for comparison with other models. There are no error bars present for the un-sampled data as the algorithm only converged to build a model in one of the three runs.

The performance of the Tomek Links data model and the un-sampled data model is seen to have a very similar format. This is due to a minimal number of data points being removed as Tomek Links. Across the three runs, a mean of $3.06 \pm 0.31\%$ of points in the majority class were removed. The small number of points removed is expected to be due to the high dimensionality of the feature space, leading to a low rate of fulfilment of the criteria in 12. There is, however, a significant increase in algorithmic convergence when compared with un-sampled data. A model was created for every instance of model creation with Tomek Links data. This suggests that there may be potential for Tomek Links to be used alongside an oversampling technique such as SMOTE.

Neighbourhood Cleaning led to a mean reduction of 14.07 ± 1.87 samples in the majority class. Whilst this is a significant improvement in reduction of the majority class data when compared to Tomek Links it is still not sufficient to work as a sole sampling technique for input to a model. This point is further demonstrated by the low value of recall when compared to other methods. Despite this, the AUCROC value is similar to that of the Near Miss and Cluster Centroid methods, despite the significant difference in the number of samples in the majority class.

The effect of ratio on recall was investigated for the Near Miss and Cluster Centroid

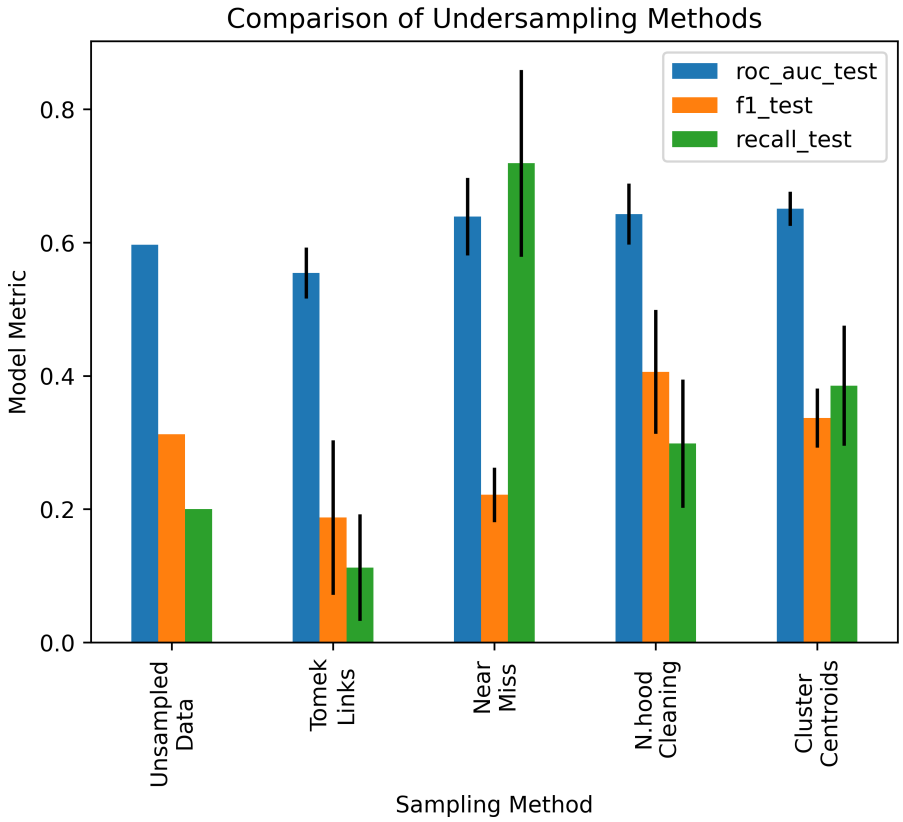


Figure 7: Bar plot to show comparison of models built using data sampled using stated under sampling techniques

methods. Figure 8 shows the influence of ratio on the recall of results. F1-score is also plotted for reference. It is shown that models created with data from the Near Miss method consistently have a higher value for recall than those created with the Cluster Centroid method. When using Near Miss, the recall rises near constantly until a ratio of 0.7 is reached. This implies that the technique is removing data points far from the decision boundary that may be causing noise in the data-set. It is hypothesised that once the ratio reaches 0.7 or higher, most outliers have been removed explaining why F1-score and recall scores remain almost constant for these ratios. This is not the case for the Cluster Centroid method, where there is an overall increase in recall with ratio coupled with an increase in error. This error is due to the random positioning of the centroids when the algorithm is instantiated. This random positioning can significantly affect results, especially when considering the WCC Survey data. This data has a high dimensionality with a limited number of data points, thus a very high number of algorithmic iterations is required for a consistent result. Overall there is a high error across all models used, demonstrating the influence of the input data on the algorithm.

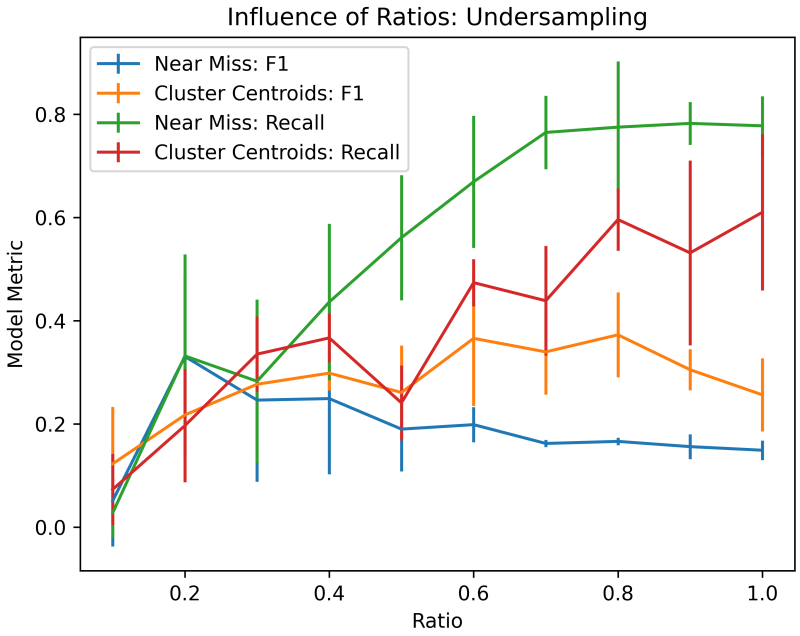


Figure 8: Plot to show the influence of changing ratio with cluster centroids and near miss under sampling techniques

SMOTE oversampling resulted in a significant positive correlation between model performance and ratio increase for ratios between 0.1 and 0.4, as shown in fig. 9. For ratios higher than 0.4 there was consistent decrease in f1 score with ratio increase. Error on each performance metric generally followed a similar pattern of peaking between 0.3 and 0.5. This is expected to be due to the random instantiating of the cluster centres at the start of the algorithm. The randomness of the points influences the lower ratios less as there is less data. For higher ratios, the higher quantity of cluster centres created will lead to a more even spread across the data set and are thus more likely to be reflective of the highly variable minority class data. A final ratio value of 0.6 was deemed to create the best, most consistent model.

Following implementation of under and oversampling techniques separately, the effect of combining them was tested. Eight new combinations were tested in attempt to show an increase in performance. These new combinations were hypothesised to have the highest potential for performance enhancement in models created. The data for these was sampled as follows:

1. SMOTE (ratio 0.6) followed by neighbourhood cleaning
2. SMOTE (ratio 0.6) followed by near miss (0.7)
3. SMOTE (ratio 0.6) followed by near miss (0.8)

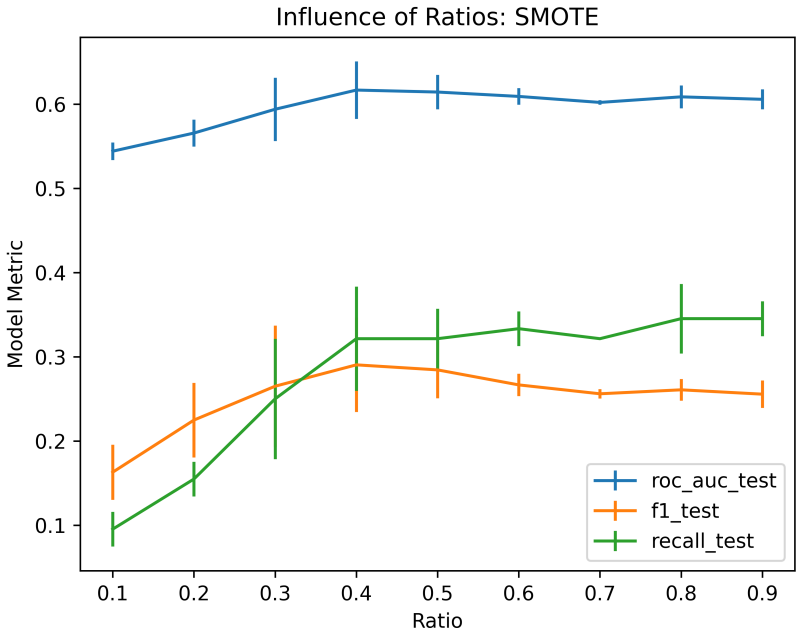


Figure 9: Plot to show the influence of changing ratio with the SMOTE oversampling technique

4. SMOTE (ratio 0.6) followed by near miss (0.9)
5. SMOTE (ratio 0.6) followed by near miss (1.0)
6. SMOTE (ratio 0.6) followed by cluster centroids (0.8)
7. SMOTE (ratio 0.6) followed by cluster centroids (0.9)
8. SMOTE (ratio 0.6) followed by cluster centroids (1.0)

Fig 10 shows the performance for each of these combinations when the model is tested on the unsampled test data. The label for each combination is given by the numeric labels detailed above. No significant improvement was seen when using models built using combination data on the test set when comparing with the models built using a data from a single sampling technique. However, significant improvement was seen across all performance metrics for models fitted on training data. This suggests a high degree of overfitting on the training data when using multiple sampling techniques.

Many different sampling techniques have been discussed with limited success. The near miss undersampling technique with an output ratio of 0.7 between the minority and majority classes was found to produce the best performing model. The models built from this data had a mean recall of 0.764 ± 0.07 and a mean F1-score of 0.162 ± 0.007. Despite being found to be the best model for the data, performance is

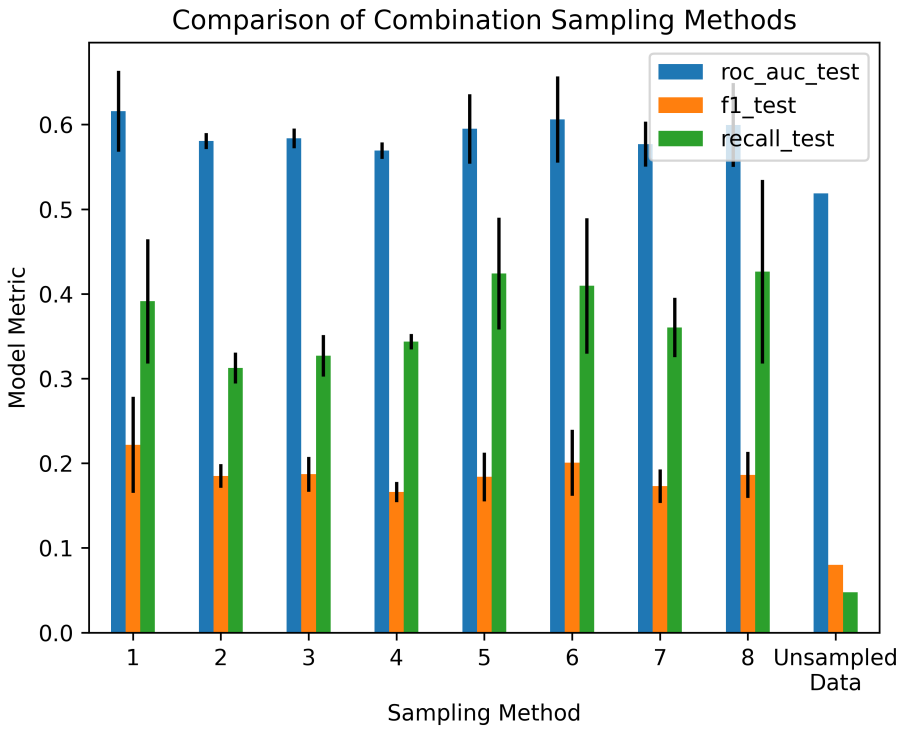


Figure 10: Bar plot to show comparison of models built using data sampled using stated under and over sampling techniques

still poor, suggesting further data cleaning or extraction is necessary. Prior to building the final model and vulnerability score, an attempt to combine the weak classifiers created from this data and other data engineering techniques will be explored is explored. It is hoped that a combination of classifiers will help boost the performance of a final overall classifier. Following this, as manual data cleaning has already been undertaken, it is suggested that the technique of implementing penalties on the data through elastic net regression should be explored.

6.4 Adaboost to improve model performance

Sampling techniques were undertaken with limited success. Thus it was deemed appropriate to further investigate methods to improve the performance of models built. Adaboost, short for adaptive boosting, is one such method. Adaboost combines a selection of weak classifiers in order to create a single stronger classifier. For this experiment, the classifiers input into the Adaboost algorithm were varying combinations of the models built from the previously sampled data. The success of models built using Adaboost was then evaluated and its potential as a technique to better evaluate the WCC Survey data was discussed.

Adaboost implements a simple boosting algorithm following input for n weak classifiers. Prior to classification, the binary target classes must be converted to have values of -1 and $+1$. For each iteration error is weighted using values from the previous iteration. Weight, W_1, is initialised as $\frac{1}{n}$. The algorithmic process is as follows:

1. For each weak classifier $\hat{h}$, calculate the error as

$$E(\hat{h}) = \sum_{i-1, y_i \neq h_i(\mathbf{x_i})}^{n} W(i) \tag{13}$$

   where y_i is the true value of the target variable for point $\mathbf{x_i}$ and $h_i(\mathbf{x_i})$ is the predicted value of the target variable for the same point

2. The classifier with the minimum error rate for the iteration, ϵ is found. The contribution of this classifier to the final model is defined by α, where

$$\alpha = \frac{1}{2} \ln \frac{(1-\epsilon)}{\epsilon} \tag{14}$$

   . The weight is then updated as

$$W_{(k+1)i} = \frac{W_{k(i)} \exp(-\alpha_k y_i \hat{h}_k(\mathbf{x_i}))}{Z_k} \tag{15}$$

   where k is the label of the current iteration

3. This process is repeated until $\epsilon_k > 0.5$ or all weak classifiers have been incorporated into the final model.

4. The final model is given by

$$\text{sign}\left(\sum \alpha_k \hat{h}_k(\mathbf{x})\right) \tag{16}$$

[Schapire and Singer, 1999].

Two independent variables were investigated for the Adaboost classification. The first was the number of models input into the algorithm and the second was the data upon which the models were built (sampled or unsampled).

In order for experimentation to be undertaken an adaboost class was created using Python 3.8 to boost and fit data. In order to utilise previously built models an adaboost function, based on the algorithm described was manually built.

Initially the WCC Survey data was split using Sklearn's *train_test_split* function [Pedregosa et al., 2011]. As stated and explored in the sampling section, machine learning models built from balanced data generally perform better than those built from imbalanced data. Using this knowledge a function was then built to divide data from the training set into n sections. This *data_split* function divided the training data into a minority class data set and a majority class data set. The majority class was then randomly divided into n subsets where n is a power of 2 (up to 2^3) as specified by the user. Each majority class subset was then appended to the minority class data set to form n new data sets. Logistic regression models were built using the n data sets and input into the *adaboost* function. Each iteration of the Adaboost algorithm was performed three times with different training and testing data splits. The mean results of each of these splits was calculated with error equal to the standard deviation.

Following this, a similar method was used for data undersampled using the near miss technique. The data input into the $data_split$ algorithm had a ratio of minority class to majority class of 0.125. This allowed for comparison between the performance of adaboost when models were created with perfectly balanced data ($n = 8$) and moderately imbalanced data ($n = 4$) and highly imbalanced data ($n = 2$).

Following the implementation of Adaboost as discussed, the output models showed no improvement on their non-boosted counterparts. This is expected to be due to Adaboost's utilisation of accuracy as a performance metric. Each weak classifier input into the Adaboost algorithm was deemed to be deceptively strong by the algorithm due to its high accuracy, which was in turn due to the imbalance in the data set. As such, W_i placed near total emphasis on whichever classifier was used in the first algorithmic iteration. Thus the contribution of the second classifier was usually two orders of magnitude lower than the first. When $n = 4$ or $n = 8$, no contribution from the third classifier was seen in 58.3% of outputs.

The utilisation of Adaboost to improve model performance was deemed unsuccessful. Further research should be undertaken on adapting the Adaboost algorithm to

incorporate recall. This would significantly improve performance with unbalanced data sets.

7 Identifying Indicators of Digital Exclusion

Section 3 provides a general background to logistic and elastic net regression. These methods are chosen due to the transparency of the models created. Logistic regression models output a series of easily identifiable coefficients which may be transformed into probability distributions for each feature input into the model. A classification model is created using elastic net regression. The coefficients of this model are used to identify the features that contribute most to the identification of an individual as $(N)DE$.

7.1 Theory

The identification and mathematical manipulation of logistic regression coefficients to output an estimated probability distribution is integral for in depth analysis of the regression output. The mathematical processes used in this report are based off methods used for categorisation of the severity of a Covid infection as discussed in [Liang et al., 2020]. They are developed slightly in order to better define confidence intervals and are discussed as follows.

The probability distribution for any given feature is given by 1. Confidence intervals indicate the likelihood this value is correct according to the variance of the input data set. Confidence intervals are calculated as

$$\left\{\mathbf{x}\hat{\beta} - z\sqrt{\mathbf{x}'\text{Var}(\hat{\beta}\mathbf{x}}\right\} \leq \mathbf{x}'\beta \leq \left\{\mathbf{x}\hat{\beta} + z\sqrt{\mathbf{x}'\text{Var}(\hat{\beta}\mathbf{x}}\right\} \tag{17}$$

where **x'** is the transpose of the feature vector and $\hat{beta}$ is the matrix of coefficients output by the regression model. z is defined as $1 - \frac{\alpha}{2}$ the parameter defined by $\mathbf{x}\beta$ can be described as being correct with the confidence interval $1 - \alpha$. Var is the variance of the feature vectors and can be found through manipulation of the co-variance matrix.

This technique is a successful indicator of confidence interval as the parameter $\mathbf{x}'\beta$ is asymptotically normal to the variance [Xu and Long, 2005].

7.2 Method

A *Vulnerability* class was built in Python 3.8, containing all functions required to build a vulnerability score. The data input into the model was the WCC survey data sampled using Near Miss under sampling at a ratio of 0.7. This was the data that produced the best performing model when investigating sampling methods. However, due to the quantity of data removed from the model it was assumed there may be

some over-fitting of the data or information loss. As such the original imputed data was also used to fit the model. Both models used the same original training and testing data splits. A mean value of the coefficients of the model was then utilised for statistical analysis.

All models produced in this experiment utilised elastic net regression using stochastic average gradient (SAGA) descent to minimise L_{log}. SAGA works as follows.

1. Create a table of $g_i = \nabla L_{log}$, where ∇ is the derivative operator, for each point, i in the data set.

2. Initialise $x^0 = g_i^0$ for each point, i, in the data set.
   Note: when representing a variable as a_c^b, b is the iteration number and c is the position in the data set

3. For each step of the model, j, pick a random point $i \in$ data set and set

$$g_{i_j}^j = \nabla L_{logi}(x^{j-1}) \tag{18}$$

   i.e. the most recent gradient of L_{logi}, keep all other values of g_i constant

4. Update the overall gradient according to

$$x^j = x^{j-1} - t_k \cdot \quad g_{i_j}^j - g_{i_j}^{j-1} + \frac{1}{n}\sum_{i=1}^{n} g_i^j \Bigg) \tag{19}$$

   where n is the number of points in the data set and t_k is the step size. [Defazio et al., 2014]

Initially the influence of hyper-parameters on the model was investigated. Hyper-parameter tuning was undertaken using Sklearn's inbuilt *GridsearchCV* function [Pedregosa et al., 2011]. This function provides functionality to iterate through all options for the hyper-parameters required in 6. However, the hyper-parameter input into the function varies slightly from 6. Input requires a value for $C = \frac{1}{\lambda}$ and $l1\_ratio = (1 - \alpha)$. C values were input on a logarithmic scale with $C \in [0.001, \ldots, 1000]$, increasing by an order of magnitude with each iteration. $l1\_ratio$ was investigated with increments of 0.1 for a range $l1\_ratio \in [0, 1]$. Following hyper-parameter tuning, the coefficients of the best model were investigated to build a vulnerability score.

Following fitting it was deemed that all coefficients implying a 5% increase or decrease in the likelihood of digital exclusion were statistically significant. All variables below this threshold were discarded. If a feature displays statistical significance close to the threshold, its contribution is discussed and evaluated for use in the final building of the vulnerability score. The results of the logistic regression models built from the data with total digital exclusion and digital exclusion due to solely owning a mobile

phone varied somewhat in their results. The following sections discuss the models built from each data set.

Errors for all values in section 7.3 are calculated using variance of individual models combined with the standard deviation of coefficient values across the two models created for each type of digital exclusion. Combination is undertaken using the following formula

$$\Delta mod_{mean} = mod_{mean} \times \sqrt{\left(\frac{\Delta mod_1}{mod_1}\right)^2 + \left(\frac{\Delta mod_2}{mod_2}\right)^2 + \left(\frac{Std_{mean}}{mod_{mean}}\right)^2} \tag{20}$$

where Δmod_x is equal to the error on the model x. mod_x is a list of coefficients created by elastic-net regression. When $x = mean$ this is implies a mean value has been taken of models 1 and 2. Std_{mean} is the standard deviation of the coefficients input into the model.

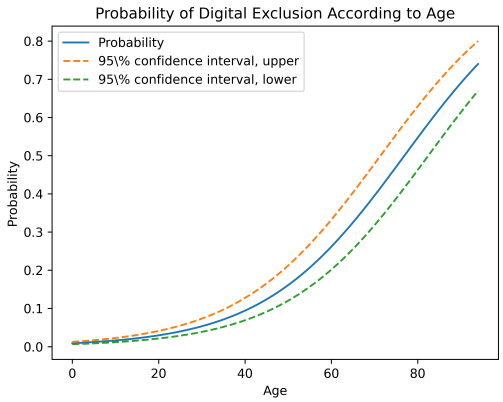
7.3 Results and Discussion

Elastic-net regression models were created to identify influential factors of digital exclusion. Both total digital exclusion and mobile digital exclusion as defined in section 5.1 were investigated. Total digital exclusion data models were trained on a data containing 726 samples. Mobile digital exclusion data models were trained on data containing 676 samples. These sample sizes are small when compared with other reports detailing use of regression models for the creation of vulnerability scores such as [Liang et al., 2020]. When analysing results this may have an influence on the stated error and as such all errors on results stated in this section should be viewed an an estimate for minimum error.

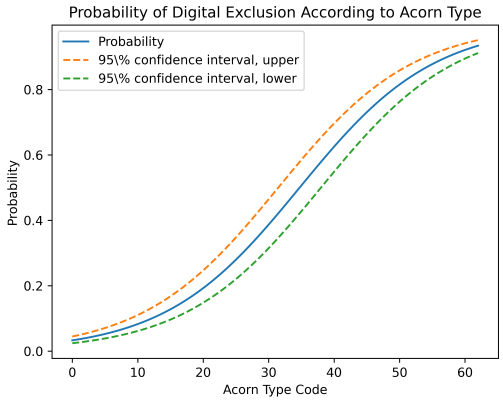
7.3.1 Indicators of Total Digital Exclusion

Of the 67 features input into the elastic net regression algorithm, the influence of 11 features was reduced to zero due to the algorithm. A further 32 features were immediately discarded due to statistical insignificance.

The two most influential factors on digital exclusion are age and Acorn type. The probability distribution of these features vs. feature value are shown in fig.11. Referencing the description of Acorn type codes found in fig. 2, it can be seen that those facing increasing financial hardship are more likely to be digitally excluded. Interestingly Acorn type codes 10-13 refer to individuals with high wealth in retirement, however, converse to later findings, there is no significant increase in the likelihood of digital exclusion for individuals in this age category. This may be because retired individuals who have significant savings are more likely to have worked higher paid jobs. These jobs may have involved working with newer technologies than their low earning peers. Furthermore, increased wealth allows for easier access to paid courses and teaching on internet use if an individual desired.



(a) Age



(b) Acorn Type Code

Figure 11: Probability distributions to show the influence on stated features on the likelihood of total digital exclusion

The probability distribution for age shows a near zero chance of digital exclusion in individuals under the age of 30. This is representative of the fact that individuals in this age bracket have grown up using computers and indeed, individuals under 29 have never experienced a world without the world wide web [Berners-Lee, 1989]. From the age of 50 the probability of digital exclusion increases sharply, with an equal chance of digital inclusion/exclusion seen at around 70 years of age. There is a further near constant increase in the probability of digital exclusion with age following this point. The reasons for this are unclear, however it is hypothesised that individuals at this age may not find a need to get online. As such, it is suggested that focus should be placed on educating older individuals on the benefits of digital inclusion.

Correlated with age, retired individuals are seen to be $20.05 \pm 8.04\%$ more likely to be digitally excluded. Retirement status is not the only job status that has influence on whether or not an individual is digitally excluded. Individuals working full time (more than 30 hours weekly) are $20.77 \pm 8.83\%$ less likely to be digitally excluded. This may be a direct result of the nature of the job an individual is undertaking. However one should also consider the age demographic of full time workers. A report from the Institute of Fiscal Studies suggested 16% of individuals aged 50-69 aim to work fewer hours, whilst for some individuals in this age bracket, part time work aids a gradual transition to retirement [Crawford et al., 2021]. This suggests that individuals working full time may be from a younger demographic and thus less likely to be digitally excluded. Finally unemployed individuals who are unemployed due to ill health are 6.59 ± 0.47 more likely to be digitally excluded. This correlates with literature discussed on the topic and may imply the necessity for more ability inclusive internet platforms such that those with additional needs are not left digitally excluded.

Perhaps correlated with unemployment due to ill health, it was found that those who felt a lack of companionship or felt isolated from others were $8.31 \pm 2.39\%$ more likely to be digitally excluded. Although the exact reason for the isolation and a lack of companionship are unknown, the mental health of those who are digitally excluded may be an interesting avenue of research to follow. Whether or not digital exclusion has an influence on an individual's feeling of isolation in an increasingly technological world or whether a lack of companionship and similar illnesses may prevent an individual from exploring new technologies are both questions that may influence further research on digital exclusion. This correlates with findings in [Doody et al., 2020], stating that older individuals who live alone are more likely to not access the internet.

Arguably the most significant and unexpected finding of the regression model was the influence of ethnic background on digital exclusion. The probability of digital exclusion was somewhat lessened for individuals from White, Western European ethic backgrounds whilst it was increased by a mean of $8.93 \pm 3.4\%$ for those of South Asian origin (including Indian, Pakistani and Bangladeshi as well as those of other Asian origins).

The final significant factor found to influence digital exclusion was the number of

council services accessed in the last year. The council services included in this measure were:

- Social Care Services (Adults or Children)
- Westminster Employment Services
- Special Educational Needs
- Homelessness Services
- Debt Advice
- Westminster Connects Shielding Support

The nature of these services suggests that individuals facing social or economic difficulties are more likely to be digitally excluded. This demonstrates that digital inclusion may still be viewed as a luxury only to be explored when an individual achieves social and economic stability. This view may be damaging to an individual's attainment of social and economic stability due to the digitally integrated nature of many basic services across the UK. As such it is suggested that opportunities for digital education should be incorporated or advertised in centres hosting the above council services.

7.3.2 Indicators of Mobile Digital Exclusion

Of the 67 features input into the elastic net regression algorithm, the influence of 9 features was reduced to zero due to the elastic net regression penalties. A further 45 features were immediately discarded due to statistical insignificance.

As with total digital exclusion, age and retirement status remained indicative of mobile digital exclusion. In addition, if an individual owned a property outright, they were $9.04 \pm 4.85\%$ more likely to suffer from mobile digital exclusion. These features seem to imply an older, more affluent group of individuals who may not own devices more than a mobile phone out of choice, rather than necessity. It is hypothesised that retired, financially-stable individual may have less need to access the internet for tasks such as job applications or online learning and as such may not see the need for investment in further electronic devices. Further surveys should be undertaken to investigate this hypothesis.

In comparison to this group of older, more affluent individuals, low financial status and economic well-being had a much higher influence on mobile digital exclusion than total digital exclusion. Those who are registered unemployed and receiving Job Seekers Allowance are $5.58 \pm 0.08\%$ more likely to suffer from mobile digital exclusion. The influence of reported financial well-being is shown in fig. 12. A score of zero correlates to a survey response of 'I am very comfortable financially,' scores increase according to level of financial worry concluding with a score of six correlating to a survey response

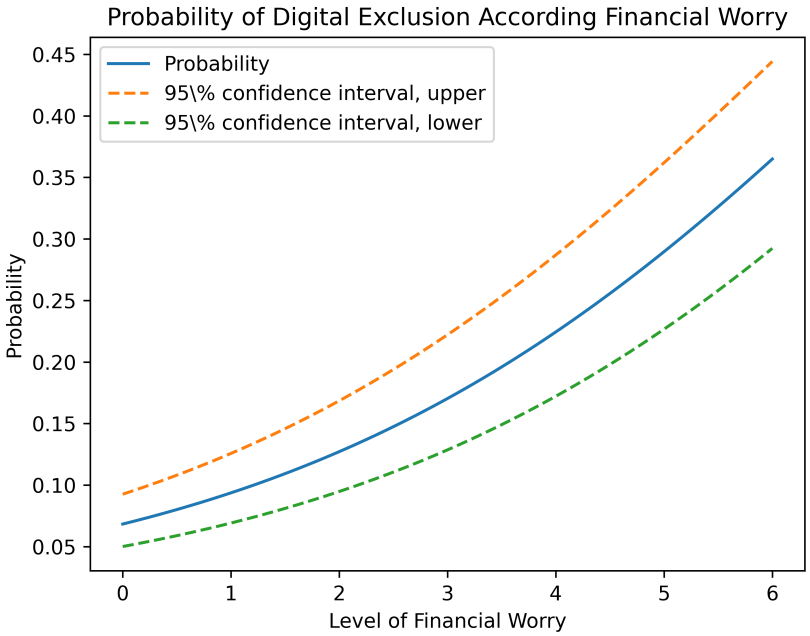


Figure 12: Probability distribution to show the influence of financial concerns on the likelihood of total digital exclusion, 0 is no financial concern with increased values implying increased financial concern.

of 'I am an not managing financially, and often have to go without essentials or am falling deeper into debt.' Although the probability of mobile digital exclusion remains low across the spectrum of reported financial well-being, there is a significant positive correlation with increased probability and mobile digital exclusion. There are two suggested reasons for these trends. The first is that individuals who are unemployed and not managing financially may view the purchase of any technological devices more than a mobile phone as an unnecessary luxury. The second is that individuals only owning a mobile phone has less access to online job applications and free teaching resources that may improve financial stability [Napoli and Obar, 2014]. Whatever the nature of the reason behind mobile digital exclusion, individuals who only have access to a mobile phone will have greater difficulties completing tasks related to job applications such as word processing. It is suggested that targets should be made to allow more universal access to electronic devices such as laptops or computers to greater aid a wider use of internet resources.

Ethnicity and citizenship played some role in mobile digital exclusion. Whilst no ethnicities indicated a statistically significant increased likelihood of mobile digital exclusion, individuals from a white, Western European ethic background were 15.47 ± 7.05 less likely to be classed as mobile digitally excluded. Furthermore UK citizens were classed as $9.88 \pm 4.32\%$ less likely to be digitally excluded. The exact reasons for

this are unclear however they imply that, along with total digital exclusion, it may be useful to situate advertisements or teaching resources about internet use in more culturally diverse locations across the borough.

Similar to total digital exclusion, whether an individual felt isolated from others or a lack of companionship implied a $6.65 \pm 0.08\%$ increased probability of mobile digital exclusion. This reiterates further suggestions of future investigations to explore the influence of mental health on digital exclusion.

It should be noted that, in general, all factors indicative of mobile digital exclusion provide significantly lower correlations with digital exclusion than their totally digitally excluded equivalents. This may be due to the decreased sample size of the mobile digital exclusion data leading to lower levels of convergence in the created regression models. However, it may be due to a higher influence of the attitudes of individuals on mobile digital exclusion. Many individuals may not see technological devices more than a mobile phone as necessary for day-to-day life. This hypothesis should be investigated in order to gain a greater understanding of why an individual only owns a mobile phone and whether it is a choice or a forced decision due to financial hardship.

8 The Vulnerability Score

A vulnerability score may prove a useful tool for identifying digitally excluded individuals across the borough of Westminster and across London. Associating demographic factors such as age and job status with a likelihood of digital exclusion may allow for easy identification of digitally excluded individuals in an area. Inputs into the vulnerability score may be available as data already collected for different purposes. This will allow for a cheap and efficient method to identify target individuals and locations for interventions to improve digital inclusion.

Following analysis of models produced by elastic-net regression in section 7.3, five indicators were chosen as decisive indicators of digital exclusion. These are age, ethnic background, job status, financial worry and feelings of isolation. For each of these features the coefficients were analysed and converted into positive or negative scores dependant on whether they increased or decreased likelihood of digital exclusion respectively. Each 5% increase in likelihood of digital exclusion increased the score attached to the value of the feature by 1. Likelihood values are rounded to the nearest 5% for the purpose of vulnerability score building. A final summation of the score values for each individual was undertaken. If an sample output a negative score, the individual was classed as NDE. If an sample output a positive score, the individual was classed as DE. Two scores were created indicative of total and mobile digital exclusion.

A table of values and their corresponding scores is shown in 1. Despite the influence of Acorn type code on total digital exclusion, it is not universally easy or free to ascertain. As such it was not included in the score. Instead small scores were associated

with increased financial worry in the total digital exclusion score in order to provide some reference to the findings reported by Acorn type code.

8.1 Performance of Vulnerability Score

The vulnerability score was tested on all survey answers to the WCC Survey. Target data was set as discussed in section 5.1. 1038 samples were available for total digital exclusion analysis and 966 samples were available for mobile digital exclusion analysis. Recall, accuracy and F1-score are used as metrics to analyse predictions made by the vulnerability score.

The metrics for the vulnerability score for total digital exclusion were 0.382 for recall, 0.310 for f1 score and 0.872 for accuracy. The metrics for the vulnerability score for mobile digital exclusion were 0.225 for recall, 0.209 for f1 score and 0.874 for accuracy. This performance is similar to early attempts at logistic regression techniques with un-sampled data. The slight increase in the recall score over the f1-score suggests that there are fewer false negatives than false positives. This is beneficial as the score aims to identify all individuals who may be digitally excluded. However, performance metrics are low for both vulnerability scores. As such this vulnerability score is deemed to be unsuitable for use in its current form. This is hypothesised to be due to the relatively small sample size of the data. As such, the methods used to create the vulnerability score may be replicated if more data becomes available.

9 A Location-Based Clustering Algorithm

The creation of hubs for individuals to be taught digital skills across Westminster is a primary goal for WCC over the coming years. The situation of these hubs is of primary importance in order to maximise the number of people able to access the centres. An initial attempt was made to incorporate location as a feature in the logistic regression model created from the WCC survey data. However, when utilising location as a feature to create the model the performance of the model decreased significantly, thus the influence of location was removed. As discussed in section 7, age, job status and level of financial concerns heavily influenced the likelihood of digital exclusion. The quantitative socio-economic data available included income, and number of individuals within a given age bracket per postcode. It was assumed that income correlates on some level to job status and financial concerns.

Alongside this, Ofcom data on broadband speeds and availability allowed for some metric to be created identifying a postcode as likely/not likely to be classed as DE. K-modes clustering was then applied to this data to segment postcodes into five clusters in the hope that each of these clusters may create a basic profile of an individual living in that cluster. These clusters can then be used alongside analysis

Feature	Value	Associated Score: Total	Associated score: Mobile
Age	Under 40	-8	-8
	40-64	0	-2
	65-74	+1	0
	75-80	+2	+2
	80 and over	+4	+3
Ethnicity	White, Western European	-4	-3
	Asian or Asian British - Pakistani	+1	0
	Asian or Asian British - Indian	+1	0
	Asian or Asian British - Bangladeshi	+3	0
Job Status	Retired	+3	+2
	Unemployed due to ill health	+1	0
	Registered Unemployed (Job Seeker's Allowance)	0	+1
	Working - Full Time (30+ HRS)	-4	-4
Level of Financial worry	0-2	-1	-1
	3-4	0	0
	5-6	+1	+1
Feelings of isolation	Hardly Ever or Never	-1	0
	Some of the time	+1	0
	Often	+2	+1

Table 1: Table of features and their corresponding values input into vulnerability score.

in section 7.3 to guide the location of the council run digital skills hubs across the borough.

9.1 Theory

K-modes clustering is an unsupervised learning algorithm that allows for the inclusion of categorical variables in the clustering technique. K-modes clustering is based off the k-means. Similarly to k-means clustering, k cluster centres are initialised randomly with the algorithm, however following initialisation, alongside the traditional Euclidean distance metric used in K-means clustering, a Boolean truth value of 0 or 1 is assigned to dissimilarities in data and added to the overall distance measure. This measure is then minimised through reallocation of cluster centres until a local optimum is reached [de Vos, 2015–2021].

9.2 Method

Three data sets were combined to provide input data into the k-mode clustering algorithm. These data sets were the WPAG, WPD and Ofcom Connectivity Data. Descriptions of these data sets are provided in section 4.2. Data cleaning and engineering was undertaken prior to input into the k modes algorithm. Following this a *location* class was built to implement k-modes clustering using the *kmodes* python package [de Vos, 2015–2021]. Five clusters were used in the final clustering algorithm. Values below five were deemed too low to provide useful distinctions between clusters risked and significantly over fitting of the data. K values above often led to a lack of convergence or over-complication of the output without any significant increase in information provided. Further functions were created for plotting and data conversion.

9.2.1 Data Cleaning and Engineering

Across the data sets, the highest total population of any postcode is 404, with a significant proportion of postcode areas having a population of 1. This disparity in postcode population was overcome through the use of COA as a metric to even out the number of residents per area.

WPAG divides individuals by gender and age brackets. Age brackets are on average five years in length from 5 years to 100+. In order to reduce the data complexity, some age brackets are combined. This combination was devised utilising the analysis in section 7.3 and existing literature. Much data from digital divide research divides individuals into generational brackets [Van Dijk, 2020]. Two significant temporal dividers include 1980 and 1993. Those born after 1980 were brought up and educated with digital media and those born after 1993 were brought up and educated with the world wide web. In developed countries, most research indicates near 100% usage for individuals under the age of forty. Further research indicates significant differences in internet use in older generations with the so called, ýoung old ́(65-75) demonstrating

a much higher proportion of internet use than those over 80 years. This existing literature correlates highly with fig 11a. which demonstrates a significant increase in the likelihood of digital exclusion at the 50. Taking into account these indicators, age data is compressed into 10 categories: Under 25, 25-39, 40-65 followed by incremental steps of five for all individuals post pension age.

WPD states absolute income for individuals at a given postcode as well as mean, median, mode and lower quartile income. The statistical indicators were input into the model and absolute values were ignored. This prevented models being fit according to outliers and other general noise in the data.

The Ofcom connectivity data supplies details of the internet connectivity speed across postcodes in Westminster and provides a later indication of digital exclusion after clustering and within each of the k-modes clusters. A binary indicator was used, dividing postcodes into two categories. If over 80% of households have an internet speed of 30 Mbitss^{-1} or over 50% of households have Gigabit availability, a postcode is classed as NDE. Any postcodes that do not reach this criteria are classed as DE.

Using Pandas dataframes in Python 3.8, a *dataset_combine* function combined data sets using postcode as a primary index [pandas development team, 2020]. As with the data input into the logistic regression model, all continuous data was normalised to prevent inappropriate model weighting towards data with larger absolute values.

9.3 Results and Discussion

Five clusters were produced using k-mode clustering. These clusters clustered postcodes according to the age and gender of the populations, the mean disposable income and the predicted digital exclusion. Mean values of selected features of each cluster were plotted against the mean value for digital exclusion. Fig 13 shows a strong negative correlation between the number of individuals aged over 65 and the rate of digital exclusion per cluster. This contradicts all previous research and findings and implies fault in the defined criteria for digital exclusion. It is concluded that the criteria referencing broadband connectivity in an area does not encapsulate the demographics of individuals within a population, rather it references the overall infrastructure of an area. This may be useful when digital exclusion references no or low facilities for internet connectivity in an area, however, the internet speeds available across Westminster are generally above the acceptable level of $10MBs-1$ as stated in Ofcom 2020a. As such digital exclusion was removed as a feature from the clustering input and focus instead shifted to mapping the age and income of individuals in Westminster. It will then be possible to manually link findings back to those referenced in section 7.3 in order to suggest good locations for the creation of internet hubs.

Figs. 15 and 16 show key attributes of each cluster created. Clusters 1, 2 and 3 in general appear to have high incomes with few individuals at retirement age. Thus it is expected that few people will be digitally excluded using both total and mobile digital exclusion definitions.

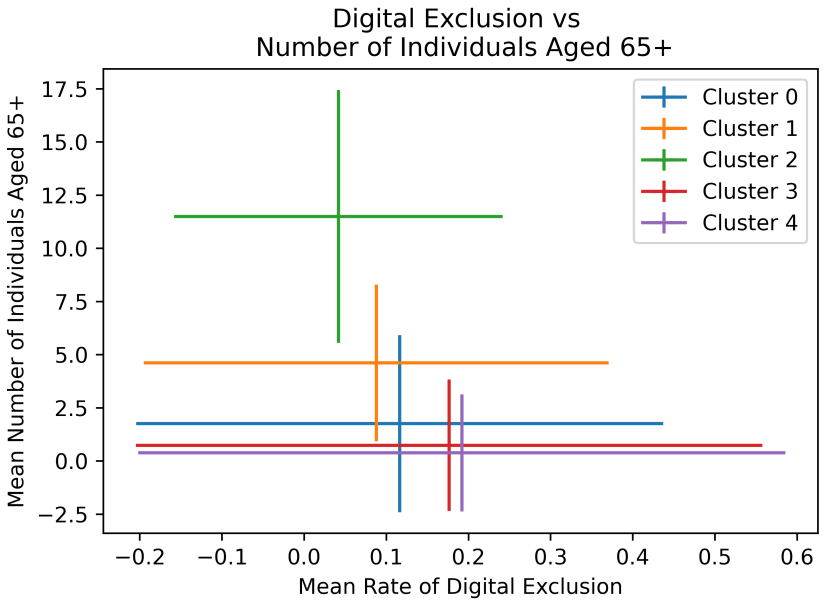


Figure 13: Plot to show the correlation between age and digital exclusion for mean values of clusters built using k-modes clustering

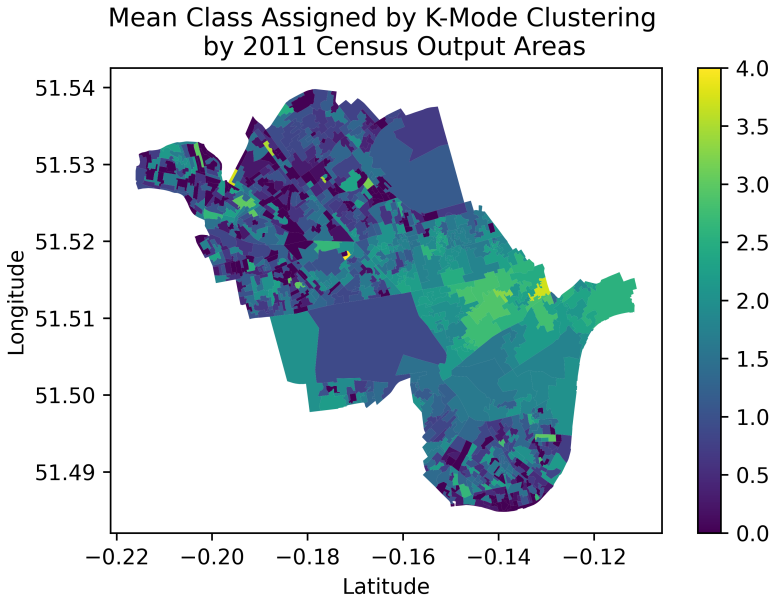


Figure 14: Map of the borough of Westminster showing the mean cluster assigned to census output area using k-modes clustering.

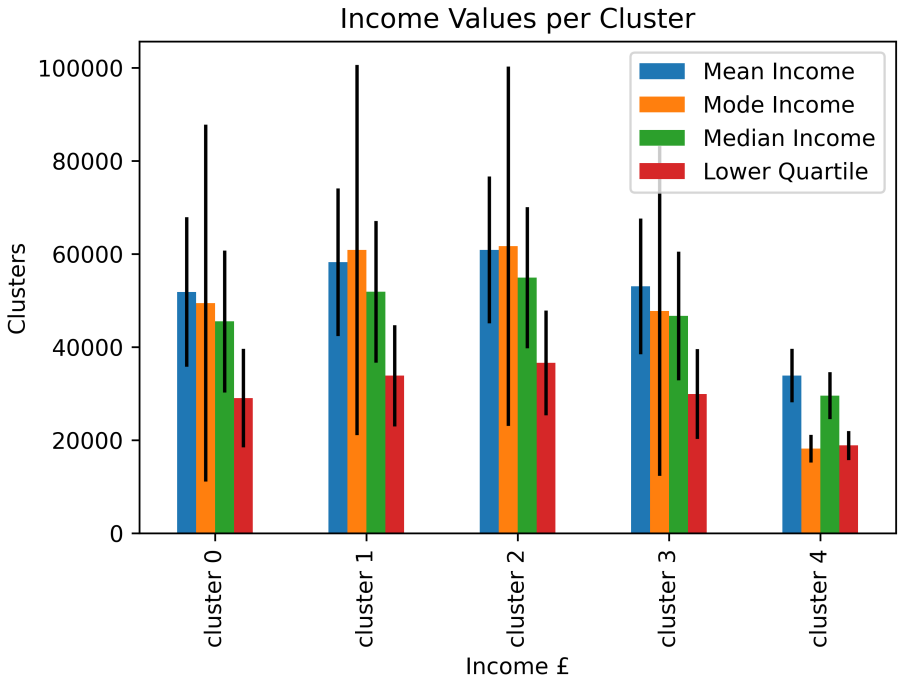


Figure 15: Bar plot to show variation of income values across each cluster identified using k-modes clustering

Cluster 4 has significantly lower income levels than the other clusters. This cluster is made up of 993 individual postcodes and is mainly situated in the east of the borough. Fig. 15 shows cluster 4 has a modal income below the London living wage of £21,157.50 per annum [MayorOfLondon, 2021]. This may be indicative of high levels of financial concerns and higher rates of unemployment. In addition, it contains few individuals over the age of 65. As such, following from discussions in section 7.3.2, it is suggested that areas in cluster 4 are targeted for interventions to prevent mobile digital exclusion.

Cluster 0 is the largest cluster and is compromised of 2188 individual postcodes. This cluster contains the highest number of individuals over the age of 65 per postcode with a mean value of 10.88 ± 5.37 individuals. Income values are similar to those in cluster 3, however lower than clusters 1 and 2. As such it is suggested that this cluster represents postcodes older individuals of stable, but not exceptional, financial status. The high influence of age and acorn type on total digital exclusion as discussed in section 7.3.1 suggests that output areas classified as cluster 0 should be targeted for interventions to prevent total digital exclusion. As seen in fig. 14 these are located in the north-west and south of the borough.

The high error in all results quoted in this section suggest that all clusters have high variance and as such the reliability of conclusions drawn should be questioned. Further investigations should be undertaken to find adaptations to clustering algorithms to

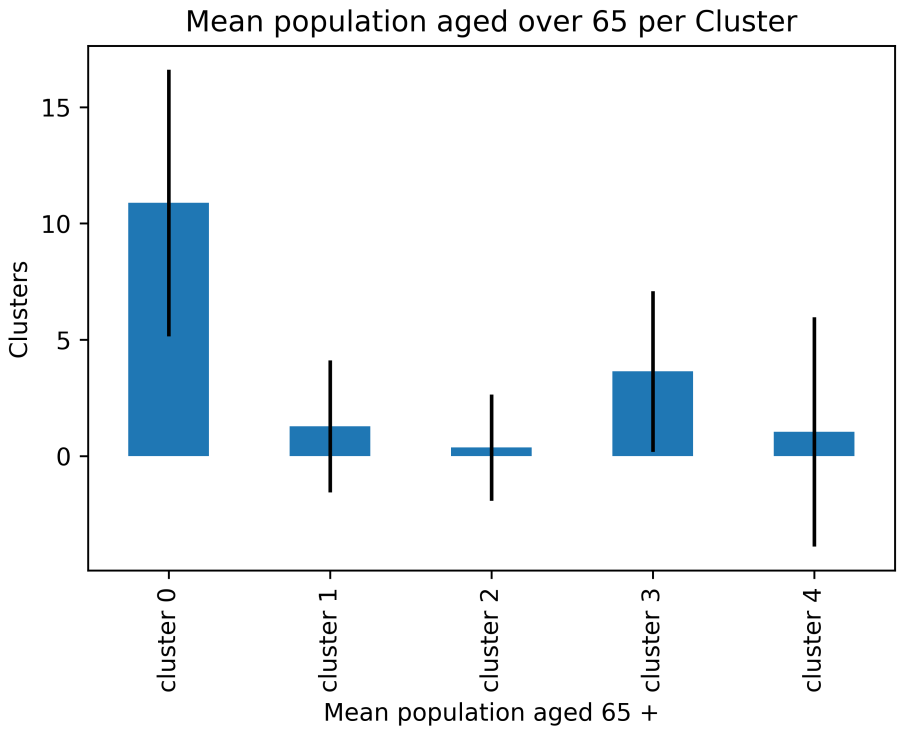


Figure 16: Bar plot to show variation of number of individuals aged over 65 across each cluster identified using k-modes clustering

increase performances in a high-dimensional and high-noise feature space.

10 Legal, Social, Ethical and Professional Issues

Data privacy statements required by Westminster City Council have been signed. The statements have been upheld for all data used in this report. To ensure security throughout the project, all work has been undertaken locally using private WiFi networks. Any code uploaded to version controlling software such as github has been uploaded to private repositories with no private data attached.

All work undertaken for this MSc project has been undertaken following the Chartered Institute for IT code of conduct [BCS, 2021] and the Institute of Engineering and Technology rules of conduct [IET, 2021]. Throughout the project all effort has been taken to keep knowledge and skills up to date with current research and publications. Information has not been represented incorrectly with regards to the performance of any tools created in this report.

The results of this report may have a social impact on due to future actions that may be taken due to this report. All results are offered as guidance and their reliability has been discussed. As such it is deemed the readers responsibility to make an informed decision on future actions. Due to limitations with data, it is not recommended that the vulnerability score created should be used as a metric to identify digital exclusion in its current form.

11 Conclusion

The two primary aims of the project were to provide WCC with guidance on the characteristics of individuals that may be digitally excluded and to provide WCC with ideas and information on how to approach the analysis of both qualitative and quantitative survey responses with machine learning. The aim of providing guidance on the characteristics of individuals that may be digitally excluded was fulfilled to some degree. Despite high errors on data due to the the data quality and quantity, it was found that age, ethnicity, feelings of isolation, level of financial concern and job status were found to be significant indicators of digital exclusion. Furthermore individuals living in the north-west, south and east of the borough were deemed most likely to be digitally excluded. It should be noted that models created may be subject to some degree of use bias due to the unintended use of the survey data for a digital exclusion classification model [Modgil, 2020].

The aim of providing WCC with information on how to approach analysis of both qualitative and quantitative survey responses with machine learning has been fulfilled. The WCC Survey data set contained limited samples, high dimensionality of the feature space, data imbalance and many missing values. As such significant research was undertaken to find the best methods of data cleaning, sampling and

imputation. The methods explored can be replicated by individuals wishing build a machine learning model and this report can provide useful guidance on how to handle more challenging data sets. Despite the relative success of some techniques, no methods used led to the creation of a very well performing machine learning model. As such it is suggested that there are many potential avenues for further research, as discussed below.

11.1 Future Research

There are many avenues of research that may be undertaken to help further identify and tackle digital exclusion across Westminster. It is hoped that research undertaken may be useful to other similar inner city boroughs. The range and focus of future research should not be understated, a few of many potential possibilities are listed below:

- **Other machine learning models** The vulnerability score created using logistic regression may prove useful when access to a computer or data limited. However, it is expected that better models may be created from WCC survey data. Logistic regression is a useful machine learning algorithm due to the transparency of the output, however, it is expected that other model creation algorithms may create better performing models. An attempt was made to test this hypothesis, however, due to time constraints a full analysis was unable to be completed. Further research should be undertaken into the creation of better models as, due to the availability of technology to WCC staff, a model would be able to classify an individual's response to a digital exclusion questionnaire in real time.

- **Further research into characteristics of digitally excluded individuals** The creation of the vulnerability score led to unexpected findings with regards to mental health and the ethnicity of digitally excluded individuals in the small data set used in this research. WCC have since created and undertaken a second survey on the attitudes and demographics of digitally excluded individuals across the borough. This survey may provide information to aid future initiatives to get people online however a survey has limited scope to understand the full extent of an individual's thought process towards the internet. It is suggested that, should time and funding allow, it may be beneficial for a more qualitative sociological approach may be beneficial in order to more fully understand why certain demographics are less likely to be online. For example, in the creation of the vulnerability score, it was found that individuals of a South Asian ethnicity were more likely to be digitally excluded. It may be beneficial to conduct longer interviews with a small number of people from this demographic in order to understand if there are any social or cultural reasons for the increased rate of digital exclusion, or whether the increased rate was merely coincidental in the given model.

- **Natural language processing for efficient data pre-processing** A large proportion of the time spent on this project was used undertaking data pre-proccesing tasks prior to the creation of the machine learning model. The majority of research in digital exclusion in Westminster is undertaken using surveys with significant quantities of categorical data. Natural language processing provides a tool to allow for greater understanding of categorical features and thus may be useful for more efficient survey analysis.

- **Ontologies for adaptive data input** An ontology is a formal description of data and knowledge defining domains and concepts as well as the relationships between them. WCC frequently carry out surveys to better understand the attitudes and demographics of people living in the borough. Many of these surveys combine multiple topics and will be used by many different departments for different purposes. The creation of a centralised ontology may allow for more efficient data extraction and manipulation in future research. In addition it will prevent the duplication of experimentation and encourage universal formatting for easy input into machine learning models. This technique has been successfully implemented across many scientific disciplines such as materials science [Ghedini and Goldbeck, 2021].

References

C. Baker, G. Hutton, L. Christie, and S. Wright. Covid-19 and the digital divide. *UK Parlimenr: POST*, 2020. URL https://post.parliament.uk/covid-19-and-the-digital-divide/. Accessed: 08/06/2021.

BCS. Bcs code of conduct. 2021. URL https://www.bcs.org/membership/become-a-member/bcs-code-of-conduct/. Accessed: 02/09/2021.

J. Berner, M. Rennemark, C. Jogréus, P. Anderberg, A. Sköldunger, M. Wahlberg, S. Elmståhl, and J. Berglund. Factors influencing internet usage in older adults (65 years and above) living in rural and urban sweden. *Health informatics journal*, 21 (3):237–249, 2015.

T. Berners-Lee. Information management: A proposal. *http://www. w3. org/History/1989/proposal. html*, 1989.

S. F. Buck. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society. Series B, Methodological*, 22(2):302–306, 1960. ISSN 0035-9246.

M. Buckland and F. Gey. The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12–19, 1994.

CACI. Acorn technical guide. 2019a. URL https://acorn.caci.co.uk/downloads/Acorn-Technical-document.pdf. Accessed: 22/04/2021.

CACI. Acorn pen portraits. 2019b. Private Correspondance.

A. Cameron, K. Rodgers, A. Ireland, R. Jamdar, and G. A. McKay. A simple tool to predict admission at the time of triage. *Emergency medicine journal : EMJ*, 32(3): 174–179, 2015. ISSN 1472-0205.

A. Cameron, A. J. Ireland, G. A. McKay, A. Stark, and D. J. Lowe. Predicting admission at triage: are nurses better than a simple objective score? *Emergency Medicine Journal*, 34(1):2–7, 2017. ISSN 1472-0205. doi: 10.1136/emermed-2014-204455. URL https://emj.bmj.com/content/34/1/2.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *The Journal of artificial intelligence research*, 16: 321–357, 2002. ISSN 1076-9757.

S. Clémençon and R. Vogel. On tree-based methods for similarity learning. In *Machine Learning, Optimization, and Data Science*, Lecture Notes in Computer Science. Springer International Publishing, Cham, 2020. ISBN 3030375986.

R. Crawford, J. Cribb, H. Karjalainen, and L. O'Brien. Changing patterns of work at older ages. 2021. URL https://ifs.org.uk/publications/15485. Accessed: 16/08/2021.

N. J. de Vos. kmodes categorical clustering library. https://github.com/nicodv/kmodes, 2015–2021.

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. 2014.

P. Doody, M. Wang, S. Scarlett, A. Hever, P. O'Mahoney, and R. A. Kenny. Internet access and use among adults aged 50 and over in ireland: Results from wave 5 of the irish longitudinal study on ageing. *The Irish Longitudinal Study on Ageing (TILDA)*, 1, 2020.

T. Fushiki. Estimation of prediction error by using k-fold cross-validation. *Statistics and computing*, 21(2):137–146, 2011. ISSN 0960-3174.

E. Ghedini and G. Goldbeck. Emmo: an ontology for applied sciences. 2021. URL https://github.com/emmo-repo/EMMO. Accessed: 02/09/2021.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

J. Harper. Coronavirus: Flexible working will be a new normal after virus. *BBC news*, 2021. URL https://www.bbc.co.uk/news/business-52765165. Accessed: 08/06/2021.

E. J. Helsper, A. J. Van Deursen, and R. Eynon. Tangible outcomes of internet use: from digital skills to tangible outcomes project report. 2015.

D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

T. IET. Rules of conduct. 2021. URL https://www.theiet.org/about/governance/rules-of-conduct/. Accessed: 02/09/2021.

T. Ivan. Two modifications of cnn. *IEEE transactions on Systems, Man and Communications, SMC*, 6:769–772, 1976.

G. James, D. Witten, T. Hastie, and R. Tibshirani. Comparison of linear regression with k-nearest neighbors. In *An introduction to statistical learning*, volume 112. Springer, 2013.

L. A. Jeni, J. F. Cohn, and F. De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction*, pages 245–251. IEEE, 2013.

I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

C. Kadushin. *Understanding social networks: Theories, concepts, and findings*. Oup Usa, 2012.

M.-h. Kim, S. Banerjee, S. M. Park, and J. Pathak. Improving risk prediction for depression via elastic net regression-results from korea national health insurance services data. In *AMIA annual symposium proceedings*, volume 2016, page 1860. American Medical Informatics Association, 2016.

W. Koehrsen. Introduction to bayesian linear regression. 2018. URL https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7. Accessed: 25/05/2021.

J. Laurikkala. Improving identification of difficult small classes by balancing class distribution. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 63–66. Springer, 2001.

G. Lemaître, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017. URL http://jmlr.org/papers/v18/16-365.html.

M. Leo, S. Sharma, and K. Maddulety. Machine learning in banking risk management: A literature review. *Risks*, 7(1):29, 2019.

W. Liang, H. Liang, L. Ou, B. Chen, A. Chen, C. Li, Y. Li, W. Guan, L. Sang, J. Lu, et al. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with covid-19. *JAMA internal medicine*, 180(8):1081–1089, 2020.

MayorOfLondon. London living wage. *London City Hall*, 2021. URL https://www.london.gov.uk/what-we-do/business-and-economy/london-living-wage. Accessed: 08/06/2021.

Mckinsey. Jobs lost, jobs gained: Workforce transitions in a time of automation.

S. Modgil. The philosophy and ethics of artificial intelligence: Algorithms. 2020.

P. M. Napoli and J. A. Obar. The emerging mobile internet underclass: A critique of mobile internet access. *The Information Society*, 30(5):323–334, 2014.

Ofcom. Connected nations update: Spring 2020. 2020a.

Ofcom. Uk home broadband performance, measurement period may 2020 – chart data (csv, 75.3 kb). 2020b. URL https://www.ofcom.org.uk/research-and-data/telecoms-research/broadband-research/broadband-speeds/may-2020-uk-home-broadband-performance. Accessed: 22/04/2021.

ONS. Census geography. URL https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography. Accessed: 25/05/2021.

T. pandas development team. pandas-dev/pandas: Pandas, Feb. 2020. URL https://doi.org/10.5281/zenodo.3509134.

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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

T. Piketty. *Capital in the twenty-first century*. Harvard University Press, 2018.

K. Podgórski. Computational genomics with r altuna akalin chapman hall/crc, 2021, xxii + 440 pages, £99.99, hardcover isbn: 978-1-4987-8185-5. *International Statistical Review*, 89(2):420–421, 2021. ISSN 0306-7734.

J. Poushter, C. Bishop, and H. Chwe. Social media use continues to rise in developing countries but plateaus across developed ones. *Pew research center*, 22:2–19, 2018.

R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999.

S. E. Thayer and S. Ray. Online communication preferences across age, gender, and duration of internet use. *CyberPsychology & Behavior*, 9(4):432–440, 2006.

UKGovernment. Coronavirus act 2020. 1, 2020. URL https://www.legislation.gov.uk/ukpga/2020/7/pdfs/ukpga_20200007_en.pdf. Accessed: 21/03/2021.

D. Ullah. The threatening problem of functional illiteracy: Revisiting education,. 2: 1–14, 12 2015.

S. Van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(1):1–67, 2011.

A. J. van Deursen and J. A. van Dijk. Internet skill levels increase, but gaps widen: A longitudinal cross-sectional analysis (2010–2013) among the dutch population. *Information, Communication & Society*, 18(7):782–797, 2015.

J. Van Dijk. *The digital divide*. John Wiley & Sons, 2020.

E. Velleman. *The Implementation of Web Accessibility Standards by Dutch Municipalities. Factors of Resistance and Support.* PhD thesis, 12 2018.

J. Xu and J. S. Long. Confidence intervals for predicted outcomes in regression models for categorical outcomes. *The Stata journal*, 5(4):537–559, 2005. ISSN 1536-867X.

L. Yu, R. Zhou, R. Chen, and K. K. Lai. Missing data preprocessing in credit classification: One-hot encoding or imputation? *Emerging markets finance trade*, pages 1–11, 2020. ISSN 1540-496X.

R. Yurchak. pgeocode. https://github.com/symerio/pgeocode, 2020.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.