Enhancing Economic Development in Deprived Regions through Machine Learning and Geospatial Modelling: A Case Study of Business Sectors and Accessibility

By

Eilidh Pike

In partial fulfilment of the
requirements for the degree of
MSc
in
Artificial Intelligence and Applications

Department of
Computer and Information Sciences

August 2023

Abstract

In the realm of social development, understanding the interplay between access to resources, deprivation, and economic activities is of paramount importance. This article presents an exploratory study that combines ensemble learning techniques with geospatial analysis to predict the distribution of various business types within deprived regions. The objective is to address the challenges of data imbalance, spatial dependencies, and hidden patterns while fostering evidence-based decision-making for policymakers and urban planners. The study employs a range of machine learning algorithms, including Random Forest Classifier, Decision Tree Classifier, Gradient Boosting Classifier, and Support Vector Machine, as baseline models to predict business distribution based on access and deprivation data. Additionally, geospatial weights matrices are used to capture spatial dependencies, while ADASYN (Adaptive Synthetic Minority Data Generation) technique is applied to mitigate data imbalance. Clustering methods are introduced to further uncover underlying data patterns. Results indicate that the application of ADASYN led to significant enhancements in model performance across various metrics, particularly in Random Forest and Decision Tree classifiers. However, combining the spatial weights matrix and clustering techniques did not produce substantial improvements in model performance. The integration of ensemble techniques with spatial weights matrix and ADASYN demonstrated notable improvements in predictive capabilities, showcasing the potential benefits of combining complementary methods. The study sheds light on the complexities of integrating multiple techniques, emphasizing the importance of method selection and compatibility testing when constructing ensemble models. While certain techniques showed promising results, others had mixed impacts. Future research could delve into the application of deep learning algorithms in geospatial analysis and explore the potential for reevaluating clustering techniques and model configurations.

Declaration

This dissertation is submitted in part fulfilment of the requirements for the degree of MSc in Artificial Intelligence and Applications of the University of Strathclyde.

I declare that this dissertation embodies the results of my work and has been composed by myself.

Following normal academic conventions, I have made due acknowledgement to the work of others.

I declare that I have sought, and received, ethics approval via the Departmental Ethics Committee as appropriate to my research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to provide copies of the dissertation, at cost, to those who may in the future request a copy of the dissertation for private study or research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to place a copy of the dissertation in a publicly available archive.

(please tick) Yes [X] No [ ]

I declare that the word count for this dissertation (excluding title page, declaration, abstract, acknowledgements, table of contents, list of illustrations, references and appendices is 10876.

I confirm that I wish this to be assessed as a Type 1 2 3 4 [5] Dissertation (please circle).

Signature: Eilidh Pike

Date: 16/08/23

Acknowledgements

I’d like to thank Sandra and Stuart for their constant encouragement and unconditional love and support during the most challenging phases of this journey. Their belief in me has been my greatest source of strength.

I’d also like to thank Aparna Asokan and Rowan Craig for their invaluable advice and support during the last 12 months.

Finally, I’d like to extend my thanks to Dr Nur Naim for her guidance, mentorship, and insightful feedback. Without their expertise, patience, and dedication, this dissertation would not have come to fruition.

Contents

1 Introduction

In the pursuit of sustainable socio-economic development in Scotland, understanding the relationship between access to resources, deprivation, and the distribution of economic activities has become a focus of the Scottish Government. Deprived regions often grapple with disparities in access to essential services and resources (Exeter et al, 2010; Popham, 2011) which can hinder local economic growth and perpetuate social inequalities (Macintyre et al, 2014). Addressing these challenges requires innovative approaches that harness the power of machine learning, geospatial analysis, and ensemble techniques to predict and analyse business distributions within such regions. Deprived areas present unique challenges that hinder the equitable distribution of economic opportunities. Despite Aberdeen having a prosperous economic environment, limited access to resources such as education, healthcare, and infrastructure, coupled with structural barriers, can create economic deserts where businesses struggle to thrive (Macintyre et al, 2014). These disparities not only impede the economic potential of these regions but also perpetuate social inequities, exacerbating the divide between privileged and marginalized populations. The problem is worsened by potentially misleading deprivation classifications (McCartney & Hoggett, 2023; Clelland et al, 2019).

This study centres its focus on a pivotal question: How can an ensemble model, combining class resampling techniques, geospatial weights matrices, and clustering, be effectively utilized to identify specific types of businesses and services lacking in regions with limited access to resources, and how does this approach contribute to understanding and addressing economic disparities? By creating a comprehensive framework that employs ensemble learning and geospatial analysis, the objective is to classify the intricate distribution of diverse business types within deprived regions. To achieve the objectives, a multi-faceted approach is adopted. Baseline machine learning models, including Random Forest Classifier, Decision Tree Classifier, Gradient Boosting Classifier, and Support Vector Machine, are employed to predict business distributions based on access and deprivation data. A geospatial weight matrix is integrated to capture spatial dependencies among data points, enhancing the models' ability to recognise spatial patterns. ADASYN and clustering techniques are also applied to create an ensemble model capable of predicting the distributions of businesses in urban and rural areas.

The integration of geospatial analysis and ensemble techniques is expected to provide a more nuanced understanding of the relationships between access, deprivation, and business distributions. This comprehensive approach aims to offer valuable insights that can guide evidence-based decision-making for both public policymakers and businesses seeking to invest or expand in deprived regions. A subsection

1.1 Background

Scotland is currently ranked as the fifteenth wealthiest country amongst OCED member countries (Scottish Government, 2016). Despite this economic standing, certain Scottish postcodes perform shockingly in deprivation rankings. The issue of deprivation has been the subject of extensive research for an extended period, with studies on the ‘Scottish effect’ dating back to the 1980s (Carstairs & Morris, 1989). These studies have consistently attributed socioeconomic factors to deprivation, such as poor health, (Ferguson et al, 2010; Carstairs, 1995) lower employment levels, higher premature mortality rates (Exeter et al, 2010; Popham, 2011) reduced income households and insecure employment, and a reduced access to services (Macintyre et al, 2014). Significant efforts have been made on a local and state level to address these factors, such as policies that aim to improve employment, health, education, skills and training and crime prevention, with varying degrees of success (Scottish Government, 2012; 2016; 2020). However, there is little research is into factors contributing to deprivation that are not solely linked to socioeconomic status. What is yet to be researched and understood is the relationship between an area’s access ranking, the influence of that on deprivation, and subsequently, the type and number of businesses present within a region.

Aberdeen ranks as Scotland's third most populated city, following the City of Glasgow and the City of Edinburgh, respectively. According to the most recent publication from the Scottish Government in 2020, Aberdeen City is predominantly classified as Urban, comprising 96.3% of its area. Conversely, 3.7% of Aberdeen City falls under the Rural category. Aberdeenshire comprises a diverse mix of rural areas, including Accessible Small Towns (14.3%), Remote Small Towns (6.5%), Accessible Rural (36.4%), and Remote Rural (12.3%). Of the total area in Aberdeenshire, 30.5% is classified as Urban. (Scottish Government, 2020).

The Scottish Government uses a tool called the SIMD for identifying areas suffering from deprivation (Scottish Government, 2016). Figure 1 in the appendix shows the SIMD16 Methodology. The English Index of Multiple Deprivation heavily influenced its development and has been continuously updated since its initialization by adding more sophisticated data sources with each iteration to enhance relevance and accuracy (Peterson & Blackburn, 2019). Appendix A shows the SIMD16 categories.

Table 1 presents the key changes in deprivation domains in Aberdeen from 2012 to 2016 from the Scottish Multiple Index of Deprivation.

Table 1: Key changes to deprivation domains in Aberdeen 2012 to 2016.

Domain	Number of zones in top 15% most deprived 2012	Number of zones in top 15% most deprived 2016	Increase/Decrease in Deprivation Ranking
Overall	22	9	< 13
Income	12	4	< 8
Employment	14	5	< 9
Health	48	17	< 31
Education, Skills and Training	34	46	12 >
Access to Services	13	16	3 >
Crime	53	57	4 >
Housing	41	74	33 >

Aberdeen has made some progress in addressing deprivation, specifically, the number of data zones classified among the top 15% of most deprived areas in Scotland. The number of zones has decreased from 22 to 9, which reflects the advancements in the Income, Employment, and Health Domains. However, the remaining four domains, namely Education, Skills and Training, Access to Services, Crime, and Housing, have increased the number of zones in the top 15% of deprived areas. These areas have deteriorated comparably from their previous assessment, which means its likely to conclude that these areas have not been the subject of policy improvements analogous to those in other areas. This information prompts an intriguing question: Despite the areas of Education, Skills, and Training, Access to Services, Crime, and Housing facing worsening conditions, how is it that the overall

SIMD ranking positions Aberdeen City in a relatively less deprived state than before? The potential explanation for the observed variations in the SIMD rankings may be attributed to the weighting assigned to each domain within the index. The Scottish Index of Multiple Deprivation (SIMD) allocates specific weights to the domains, influencing their overall contribution to the final deprivation rankings.

Upon analyzing the weightings, it becomes apparent that the Income and Employment domains collectively carry a substantial weight of 56%, while the remaining five domains account for a combined weight of 44%. This distribution results in a slight imbalance within the SIMD's assessment framework. It is possible that the domains showing an increase in deprivation – namely Education, Skills & Training, Access to Services, Crime, and Housing – may not be receiving adequate attention due to this disparity in weighting. The higher weighting assigned to the Income and Employment domains may overshadow the pressing issues faced by these other domains, leading to a potential underestimation of deprivation in these areas. Indices such as the SIMD have been criticized for applicability to rural areas, which brings into question the usage of data zones as a measurement for deprivation in an area, further contributing to the potential oversight of the five domains.

The SIMD has received criticism for how it measures deprivation, in terms of the weightings it gives to domains, and concerns regarding their applicability to rural areas. McCartney & Hoggett (2023) found that the actual number of deprived individuals varies widely across different regions, and Clelland et al (2019) found that the SIMD misses a higher percentage of income and employment-deprived people in remote, rural and island areas and that the absolute number of people is missed in urban areas due to temporary accommodation and secondary addresses, although this issue wasn’t strongly related to rurality. This suggests that the SIMD may have an inherent bias against rural areas, therefore, the current weightings given to the seven domains in measuring deprivation should be treated with caution. Aberdeen has issues of spatial heterogeneity, suggesting that the increase in deprivation in five domains may be linked to the bias in weightings.

Based on the evidence produced by McCartney & Hoggett and Cleland et al, the use of data zones in order to measure how deprived an area is brought into question. ‘Data zones’ are the statistical geographic unit used in the SIMD 2016. The data zones are intended to be used as a sample of the population containing roughly 500 and 1000 households. Ralston (2014) criticized this method of measurement stating that “Ward areas have a mean population of around 5500, whereas the SIMD data zones have an average population of around 780. Still, the applicability of the SIMD to rural regions necessitates critical examination and evaluation to ensure its effectiveness and validity in capturing the nuances and specific challenges these areas face. This can lead to artificial groupings of areas that might not share similar deprivation characteristics.

The domains in which Aberdeen is failing to improve in the SIMD are a cause for concern. Improving Education, Skills, and Training and Access to Services can prevent the collapse of businesses within the region. By increasing access to skill development programs, apprenticeships, and utilising employment opportunities within the area, the opportunity for stable employment and a subsequent skilled workforce results in higher incomes and an overall enhanced quality of life. Furthermore, leveraging data from Aberdeen can predict sectors of Scottish businesses likely to exist in deprived areas. Employing machine learning can aid in identifying areas requiring additional funding to improve deprivation levels. By enhancing the resilience of businesses in Aberdeen, and therefore, an increase in overall GBP, this approach can prevent business dissolution and foster growth in the region, as well as reduce overall deprivation in the area.

There is a disproportional relationship in Aberdeen concerning its high number of registered businesses and deprivation in the region. Aberdeen currently has the third highest number of companies per region at 5.9% of all registered businesses, coming behind City of Edinburgh at 17.9% and Glasgow City at 19.5% respectively (Enterprises in Aberdeen and Aberdeenshire by Division, 2020). Yet, its results from the most recent Scottish Index of Multiple Deprivation 2016 highlight missed areas of deprivation. Aberdeen ranks as the third most populated city in Scotland, following

City of Glasgow and City of Edinburgh, respectively. According to the most recent publication from the Scottish Government in 2020, Aberdeen City is predominantly classified as Urban, comprising 96.3% of its area. Conversely, 3.7% of Aberdeen City falls under the Rural category. Aberdeenshire comprises a diverse mix of rural areas, including Accessible Small Towns (14.3%), Remote Small Towns (6.5%), Accessible Rural (36.4%), and Remote Rural (12.3%). Of the total area in Aberdeenshire, 30.5% is classified as Urban. (The Scottish Government, 2022). These distinctive attributes, characterized by a mixture of rural and urban areas, and the prevalence of high business activity in the region, make it a suitable subject for testing the robustness of the SIMD in correctly identifying deprivation as a whole.

Like other Scottish regions, policy-decision making in Aberdeen is made by the Scottish Government, which relies heavily on yearly publications. However, with the emerging strong developments in machine learning, there is an opportunity to enhance this decision-making process. By utilizing data from Aberdeen, it is possible to predict which sectors of Scottish businesses are likely to exist in areas of deprivation. Machine learning can predict deprivation levels to identify areas where more funding is necessary and to improve the areas in which deprivation is failing Aberdeen by utilizing the resilience of businesses in Aberdeen, and preventing business dissolution. Machine learning can ultimately improve the overall produced revenue of Scotland and lead to a reduction in the budget deficit. The paper aims to apply this on a wider scale to benefit Scotland’s overall financial health.

Establishing a context in the presence of the SIMD requires geospatial analysis. The SIMD uses data zones as a context measure, but the work of McCartney & Hoggett and Cleland et al brings this into disrepute. Further examination revealed that there was a multitude of evidence against data zones being used in deprivation analysis. Inequality within data zones can lead to inaccurate representations of socioeconomic status and deprivation levels since data zones may not account for internal variations. Additionally, data zones may not capture rapid changes in deprivation over time, resulting in outdated measurements. Moreover, data availability might be limited at the data zone level compared to widely used and recognized postcodes, affecting data comparability and policy implications. Machine learning allows us to to measure deprivation more effectively by studying postcodes because of their level of granularity and widespread use. Combining data from both data zones and postcodes, where available, can provide a more comprehensive understanding of deprivation patterns for better policy decisions and resource allocation.

Quantitative methods of measuring deprivation take a deductive approach, where the hypothesis from theories is tested on data. However, this method is restricting and limits the information that we can gleam from data. By rejecting the idea that socio-economic theories offer sometimes true statements that only hold under specific conditions for specific individuals and accepting the idea that different mechanisms come into play, such as population heterogeneity, (Goldberg, 2011). Machine learning advancements allow us to test this theory.

Addressing the existing imbalance by using a different method of calculating deprivation can lead to a more nuanced and representative evaluation of deprivation across different aspects of societal well-being. By utilizing machine learning advancements to create a new, fairer way of weighting deprivation and shifting the focus from socioeconomic factors, the SIMD can better guide targeted interventions and policy initiatives to address deprivation in all its multifaceted dimensions effectively. Current measures for identifying deprivation are lacking at best; resources could be allocated much better, and machine learning can help to pinpoint clusters and hotspots where businesses, where skills can be learned, are, therefore improving income, and ultimately improving deprivation.

In summary, addressing deprivation requires innovative thinking and a willingness to challenge existing methodologies. By harnessing the power of machine learning and refining our approach to deprivation measurement, Scotland can make significant strides in reducing inequalities, fostering economic growth, and enhancing the overall well-being of its diverse communities.

1.2 Related work

Incorporating machine learning into the context of social development is not a novel concept, however, recent developments in computational capabilities made it possible not only to use near real-time data to identify emerging trends and anticipate future problem areas but to provide a contextual backdrop against these issues. The incorporation of geospatial data and clustering further allows the advancement of machine learning models to learn from data and predict future trends and move away from a linear framework.

Anwaar et al. (2016) illustrated this idea by examining how big data and geospatial analysis can aid in humanitarian emergencies. In the midst of the ongoing migration crisis in Syria, aid organizations encountered difficulties in efficiently distributing resources to the most affected regions and coordinating efforts with overlapping groups. To tackle this issue, the United Nations Refugee Agency (UNHCR) partnered with volunteer organizations to create a real-time heatmap named Services Advisor. This heatmap utilized information from the UNHCR's Activity Info, as outlined by Kshetri (2014), and enabled individuals to access crucial details about nearby organizations, their functions, operational hours, and capacity. By zooming in on their immediate location, migrants could receive prompt assistance, reduce extended wait times, and facilitate a more efficient strategy for addressing migration's effects using geospatial analysis. Through the utilization of spatial analysis techniques, a deeper comprehension of spatial patterns and relationships within SIMD data can be achieved, unveiling concealed trends, particularly in regions with spatial variations. This, in turn, allows for well-informed decisions regarding resource distribution, policy interventions, and targeted services to alleviate deprivation in specific geographic zones. Through the development of an expansive classification system on a real-time geospatial map, they successfully circumvented the constraints posed by conventional statistical methods. This innovation enabled the strategic allocation of resources to areas with the highest demand, and in essence, provided a step-by-step manual for efficient resource distribution. This approach could be expanded to influence policy choices undertaken by the Scottish Government, utilizing innovative techniques that ensure overlooked areas receive the necessary assistance they require.

Previous literature has used geospatial models and classification in the context of indices of multiple deprivation. Arribas-Bel et al (2017) applied geospatial analysis to Random Forests and Gradient Boost to the English Multiple Index of Deprivation. The study aimed to understand the impact of adding a geospatial weights matrix to the baseline models and how it affected their performance in predicting deprivation levels. The authors found that after incorporating a geospatial weights matrix into the models, both Random Forest and Gradient Boosting exhibited improved performance. The performance metrics were quantified using an accuracy score, where the baseline models achieved scores of 0.39 for Random Forest and 0.41 for Gradient Boosting. However, with the addition of the geospatial weight matrix, the performance scores increased to 0.54 for Random Forest and 0.5 for Gradient Boosting. This improvement suggests that incorporating geospatial information and accounting for spatial dependencies among data points contributed to better predictions of deprivation levels in the context of the English Multiple Index of Deprivation. The study's findings highlight the significance of considering spatial relationships when analyzing and modelling socioeconomic indicators. Arribas-Bel centred its research on Liverpool as a case study, which shares similar deprivation statistics to Aberdeen (EIMD, 2016) but without the blend of urban and rural elements found in Aberdeen. Therefore, this data allowed the researchers to explore the impact of geospatial analysis more effectively, isolating the effect of spatial dependencies. It’s worth investigating if another element of an ensemble model can aid this possible hurdle. An additional preprocessing technique that would allow for better classification of businesses in deprived areas would allow the model to track economic activity more accurately.

Data imbalance significantly undermines the generalization performance of supervised learning algorithms, particularly in multiclass data with geospatial elements (Kaur, Pannu, & Malhi, 2019). Synthetic oversampling, a widely employed approach, aims to rectify this issue by generating synthetic examples for minority classes, thereby balancing the class distribution. Besson et al. (2022) applied

ADASYN to rebalance datasets related to ecological monitoring and species classification, resulting in a substantial 39% improvement in the accuracy of rare species identification, supporting conservation efforts. Similarly, Pradhan & Sameen (2019) showcased the effectiveness of ADASYN in classifying road traffic accident severities in US regions, achieving a significant 32% accuracy enhancement compared to a 24% improvement with SMOTE. The concept of synthetically increasing populations can be applied to the research objective of aiding businesses in Aberdeen to stay afloat and monitor which businesses are most in need of support or extra resources.

Similarly, cluster analysis can help characterize population heterogeneity, enabling the assignment of cases to specific groups. The use of clustering has previously been explored in sociology to discover subgroups in populations and then link the emergence of these subgroups to external factors. This has reversed the traditional process of proposing a socio-economic theory and then finding a subgroup to fit it; the subgroups themselves provide a more specific, tailored theory (Bonikowski, 2016; Baumer et al., 2017). In the context of incorporating machine learning and geospatial analysis into social development, recent advancements have shown the evolution of traditional machine learning models through ensemble learning techniques. Mueller et al. (2018) demonstrated a cluster-based machine learning ensemble approach that builds upon traditional classification algorithms such as Support Vector Machines and Decision Trees and performs well on datasets with significant variance. This approach, aimed at improving model performance, aligns with the exploration of clustering as a preprocessing technique by Trivedi, Pardos, Sarkozy, and Heffernan (2018). They divided data into homogenous regions, as exemplified by Trivedi et al., who utilized clusters to enhance modelling performance in assessing student achievement. The insights gained from clustering allowed these homogeneous student groups to outperform global models significantly. This concept of leveraging clusters to extract previously unseen information also applies to geospatial data. Mueller et al.'s workflow not only outperformed independent base learners and comparative ensemble methods but also preserved local inferential capabilities, maintaining interpretable relative importance values and non-transformed coefficients. This evidence of the effectiveness of ensemble learning, combined with the insights gained from clustering, offers a novel approach to understanding various factors influencing social development. For instance, Muller et al.'s findings revealed that the significance of education and labour characteristics in estimating health insurance coverage varies between urban and rural areas, indicating the diverse impact of job types and insurance options. This emphasizes the potential of clustering to uncover distinct patterns in rural and urban contexts. By incorporating such techniques into the final algorithm, it becomes possible to address multiple research questions simultaneously, thereby enhancing the accuracy and depth of insights in the realm of social development.

In summary, the integration of machine learning and geospatial analysis into the realm of social development signifies a pivotal evolution in understanding complex socio-economic dynamics. Recent strides in computational capabilities have empowered these methodologies to not only capture real-time trends but also provide a contextual foundation for addressing emerging challenges by being able to deal with the issue of population homogeneity. The innovative use of geospatial data, clustering, and ensemble learning techniques has redefined traditional approaches, enabling predictive insights and informed decision-making.

1.3 Problem Statement

The central objective is to develop an ensemble model capable of precisely identifying the types of businesses and services that are lacking in specific regions, particularly those suffering from limited access to resources and services. To address this, the study will employ a sophisticated approach involving class resampling techniques to balance imbalanced datasets, geospatial weights matrices to incorporate spatial dependencies, and clustering to uncover hidden patterns within the data. The practical application of the ensemble model entails analyzing business data to predict the economic sectors that are likely to exist in areas grappling with access deprivation. By understanding the relationship between economic activity and accessibility, the study aims to provide insights into potential solutions for addressing economic disparities across regions.

2 Methodology

This section delves into the methodology employed in this research project, providing an in-depth exploration of the approaches, techniques, and tools used to achieve the project's objectives. The methodology details various stages, from data preprocessing and manipulation to the application of advanced machine learning models and spatial analysis techniques. Libraries, data preprocessing, model selection, optimization, and evaluation strategies are comprehensively detailed to shed light on the comprehensive framework utilized to derive insights into business distribution within deprived regions.

2.1 Data

Dataset 1 (DS1 from now on), contains an analysis of the Scottish Index of Multiple Deprivation (SIMD) in Aberdeen City, which was published by the Scottish Government on 31st August 2016. The dataset comprises 59 columns that detail factors contributing to the overall deprivation ranking of Aberdonian postcodes. SIMD16 is the sixth edition of the official tool used by the Scottish Government since 2004 to identify concentrations of deprivation in Scotland. This index combines seven distinct aspects of deprivation including income, employment, health, education, geographic access to services, crime, and housing. Each aspect is evaluated using various indicators, resulting in a ranking for each postcode. These ranks are on a scale from 1 (most deprived) to 6,976 (least deprived). Figure 1 shows the full SIMD 2016 methodology and the association of each column in each domain.

Dataset 2 (DF2 from now on), contains information regarding the type and postcode of businesses in Aberdeen City. The data was obtained from the Aberdeen City Council website (Aberdeen City Council, 2020*).*

2.2 Libraries and Environment

Pandas (Version 1.3.3), Numpy (1.21.2), Seaborn (0.11.2), and Matplotlib (3.4.3) were used for data manipulation, analysis, and visualization. Data preprocessing employed KNNImputer (0.24.2), ADASYN (0.8.1), DBSCAN Clustering (1.2.1) LabelEncoder (0.24.2), and MinMax Scaler (0.24.2). The spatial analysis utilized Geopandas (0.10.2), Libpysal (4.4.0), and Geopy (2.1.0). Scikit-learn (0.24.2) facilitated machine learning modelling, with classifiers including Random Forest, Gradient Boosting, Decision Tree, and Support Vector Machines. The model evaluation used Scikit-learn (0.24.2) metrics. Data processing and modelling occurred using Python3 and Jupyter Notebook.

2.3 Preprocessing

This section presents a comprehensive approach to enhance the quality of the two datasets. The primary goal was to refine the datasets for subsequent analysis, ensuring data integrity and optimal performance of machine learning algorithms. This section outlines the steps taken to address redundant columns, handle missing values, standardize variables, and streamline categorical features.

Dropping redundant variables. Redundant variables were identified and removed from DF1. There are 5 variables describing a postcode’s overall deprivation ranking – ‘rank’, ‘percentile’, ‘vigintile’, ‘decile’, and ‘quintile’. It was decided to use the ‘percentile’ as a deprivation ranking measure instead of ‘rank’ as the data ranged from 1-6,967 which would have caused issues when scaling the rest of the data. The ‘percentile’ range ranks the postcode area from 1 (least deprived) to 100 (most deprived). This method of measurement was chosen as it had a far more standardised measure. Some variables contained repeated information. For instance, 'urname' and 'urclass' had the same data encoded in text and disordered numbers, respectively. Consequently, 'urclass' was dropped, and ‘urname’ was retained to be encoded correctly. Similarly, 'laname' and 'council_area' held identical data, leading to the removal of 'laname'. Furthermore, the variables 'population' and 'total_population' contained the same data, and 'total_population' was retained while 'working_age_population' appeared twice, with one instance labelled as 'working_age_population.1'. In this case, 'working_age_population' was chosen to be retained. Other redundant columns were identified, including ‘dz’, ‘dzname’, 'izname', 'hbname', 'mmwname', and 'spcname', which either had outdated data or were not necessary for the analysis. These columns were dropped from the dataset. Additionally, four variables,’income’, 'crime', 'overcrowding', and 'no_central_heat', were duplicated, providing both the exact number ('count') and the percentage of the recorded population ('rate'). Given that the majority of DF1 was recorded in ratio format, and the ‘percentile’ rank was used, the decision was made to preserve the 'rate' variable of these columns. This choice aimed to maintain data integrity, ensure more precise and accurate results during feature scaling and enhance the performance of machine learning algorithms. By retaining the 'count' variable, the risk of information loss was mitigated, and consistency across the dataset was upheld. Appendix A1 shows DF1 and DF2 after dropping redundant columns.

Label encoding. Label encoding was selected over one-hot encoding to avoid significantly increasing the dimensionality of the dataset. ‘urname’ from DF1 and ‘DESCRIP’ from DF2 were both label-encoded. The resulting new columns were appended to their respective and the orginal columns were removed. Additionally at this stage, steps were taken to simplify the data representation, these instances were merged with the 'large urban areas' class.

Using K- Nearest Neighbours to fill null values. K-Nearest Neighbours (KNN) imputation was employed to fill the null values in three columns of DF1. Specifically, the 'attendance' column had 1285 null values, the 'attainment' column had 713 null values, and the 'crime count' column had 324 null values. Due to the fact that DF1 contained both continuous and catergorical columns, KNN imputation was used to handle these missing values (Scikit-Learn, 2018). KNN imputation is a method used to predict missing values in a dataset. To achieve this, the algorithm calculates the mean value from a specified number of nearest neighbours (n-neighbours). It finds these neighbours by identifying the samples in the training set that is closest to the records with missing values. These nearby points contribute to the prediction by averaging their values, resulting in a predicted value for the missing entry (Jonnsson & Wohin, 2002; Scikit-Learn, 2018). The data was scaled using an ordinal scaling method prior to imputation, then the data frame was converted back to the original format after the imputation.

Recatergorising business classes. The ‘DESCRIP’ category in DF2 is a column describing the type of business associated with a postcode. The column originally contained 232 unique values, with very imbalanced classes. These classes contained many repeating classes, for example, ‘store’ and ‘store etc’, as well as misspelled or repeated classes, for example ‘storessss’. It was considered to use SIC codes, however, the final combined two data frames would have been too small, and the machine learning models would have been too sensitive to the severe diversion of the classes. Therefore, the decision was made to classify the businesses by hand in a way that would yield the best results from the machine learning models.

Urban and Rural recatergorisation. ‘The 'Urname' column encompassed 5828 instances categorized as 'Large Urban Areas', 565 instances as 'Accessible Small Towns', 446 instances as 'Accessible Rural', and 3 instances as 'Other Urban Areas'. To address the potential adverse impact of class imbalance on the models, a strategic approach was adopted. Specifically, the labels 'Other Urban Areas' and 'Large Urban Areas' were regrouped under the 'Urban' classification, while 'Accessible Small Towns' and 'Accessible Rural' were consolidated into the 'Rural' classification. As a result of this rebalancing, the distribution of instances transformed into 5831 instances classified as 'Urban' and 1011 instances classified as 'Rural'.

Scaling numerical data using Min-Max Scaling. The variables associated with the health ranking domain variable in DF1 (‘cif’, ‘alcohol’, ‘drug’, ‘smr’, ‘depress’, ‘lbwt’, ‘emerg’,) were recorded in a standardised ratio format. In order to mitigate the effects of these values potentially offsetting the other variable values that were recorded in continuous ranges, min-max scaling was utilised so that the remaining features could be compared on the same scale. Min-Max scaling uses the following formula to transform each data point in a feature to a scaled value between 0 and 1.

This transformation aligns data to a uniform range, rendering it particularly applicable for classification models like Random Forest, Gradient Boost, and Support Vector Machines. Additionally, it stands as the preferred scaling method for clustering tasks. This practice is commonly employed in machine learning, especially when working with variables in standardized ratio formats.The application of min-max scaling was extended across all variables within DF1. This uniform scaling enabled alignment among variables to a shared scale. The result is enhanced consistency and integrity across the dataset, enabling health variables to exert an equitable influence on features and business-related focal points.

Concating the 'PCOUT' & 'PCIN' variables in DF2. The 'PCOUT' (outward postcode) and 'PCIN' (inward postcode) were combined into one column.

Concating the preprocessed dataframes. DF1 and DF2 were concated on the postcode variable so that each postcode now had the number and type of business associated with it. Postcodes that did not have any businesses associated with them were dropped from the final dataframe.

2.4 Models

Following the existing literature into examining Indices of Multiple Deprivation and Geospatial Analysis, standard methods such as Random Forest Classifier, Decision Tree Classifier, Gradient Boosting Classifier, and Support Vector Machine were utilized to produce an analysis of the impact of the Access Domain on the number and type of businesses within an area. However, unlike most studies in the literature, the traditional classifiers and regressors will be expanded using geospatial data and combined the results from the classifiers and regressors combined into an ensemble model to boost the results of the baseline machine learning models with added geospatial context. Extending traditional models to incorporate more modern machine learning techniques and applying geospatial context to create an ensemble model will hopefully mitigate the limitations of traditional models and leverage the power of more complex algorithms in order to capture non-linear relationships, such as the impact of the access domain ranking on the number and type of businesses with the area.

Decision Tree Classifier

The Decision Tree Classifier (Quinlan,1986) is a supervised machine learning algorithm that partitions the data into subsets depending on the different values, with the goal of creating tree-like structures that can make predictions on new, unseen data. The subsets are created through selected features that the algorithm believes are the most informative in predicting the target class. The features are chosen through two values: Gini Impurity and Entropy. Gini Impurity measures the probability of a randomly selected element being misclassified. Entropy measures the level of uncertainty in the data. Using these two values, the Decision Tree Classifier splits the data split the data into subsets with higher purity. This process continues until the tree can either no longer be split or criteria set by parameters are met, forming a tree that can make predictions for new data by traversing from the root node to a leaf node based on feature values. Decision trees are often used in conjunction with ensemble methods to avoid overfitting (Sklearn, 2022; Géron, A. 2023).

Random Forest Classifier

The Random Forest Classifier (Breiman, 2001) is an ensemble machine learning algorithm that enhances the predictive accuracy and robustness of the Decision Tree Classifier. It employs a technique known as bootstrapping, which creates multiple random subsets of the training dataset. For each bootstrap sample, a decision tree is then constructed. Each decision tree is then built using a different subset of the data and will have differing classification rules, resulting in different predictions. At each node in the decision tree, a subset of features is randomly selected from the total set of features. This aims to introduce a layer of randomness that makes it resilient against datasets that have a high number of correlational features, mitigating issues of overfitting. Predictions are made using a voting feature and the class with the most votes becomes the final prediction (Sklearn, 2022; Géron, A. 2023).

Gradient Boosting Classifier

The Gradient Boosting Classifier (Friedman, 2002) is an ensemble method for supervised learning tasks. It starts with an initial prediction and fits the weakest-performing models from decision trees or support vector machines to correct the errors in predictions by assessing the prior Gini Impurity and Entropy values. Data points are assigned a weight that reflects how well the previous the current model, either the SVM or the Random Forest predicts it. Data points that have a higher rate of error are assigned more weighting. The predictions of all weak learners are combined to update the overall prediction of the model. The learning rate parameter can be adjusted to determine the contribution of each weak learner to the final prediction.

Support Vector Machine Classifier

The Support Vector Machine Classifier (Cortes & Vapnik,1995) is a supervised machine learning machine algorithm that attempts to find an optimal hyperplane that best separates data points from different classes in a high-dimensional feature dataset. As the dataset contains 40 variables, SVM was selected as one of the baseline models. The SVM algorithm takes each data input and represents the feature in a multi-dimensional space. Each vector contains attributes that discuss the associated variables of the data input. The SVM aims to find a hyperplane that maximises the margin between different classes. the margin is the distance between the hyperplane and the nearest data points of each class. The nearest data points are support vectors. The goal of the SVM is to find the hyperplane that not only separates the classes but also maximises the margin. Due to the high dimensionality of the dataframe, the support vector machine classifier was also considered for a baseline model (Sklearn, 2022; Géron, A. 2023).

2.5 Model Optimization: Class Imbalances

2.5.1 Class imbalances

The 'Urname' and 'Descrip' variables were highly imbalanced. To address this imbalance, the ADASYN approach was employed as it proved to significantly improve the algorithms in the geospatial studies of Pradhan & Sameen (2019) and Besson et al, 2022. ADASYN serves as an effective countermeasure against class imbalance, particularly in scenarios involving multiclass imbalance. Unlike SMOTE, which generates synthetic samples uniformly across the entire dataset, ASDYN takes a more adaptive approach by pinpointing specific areas within the feature space where the minority class lacks representation and generates synthetic samples exclusively within these identified regions. This adaptiveness results in synthetic samples that not only bolster diversity but also accurately capture the distribution of the minority class. This approach contributes to a more realistic and representative dataset, which in turn can lead to improved performance of machine learning models trained on imbalanced data.

2.5.2 Spatial Weights Matrix

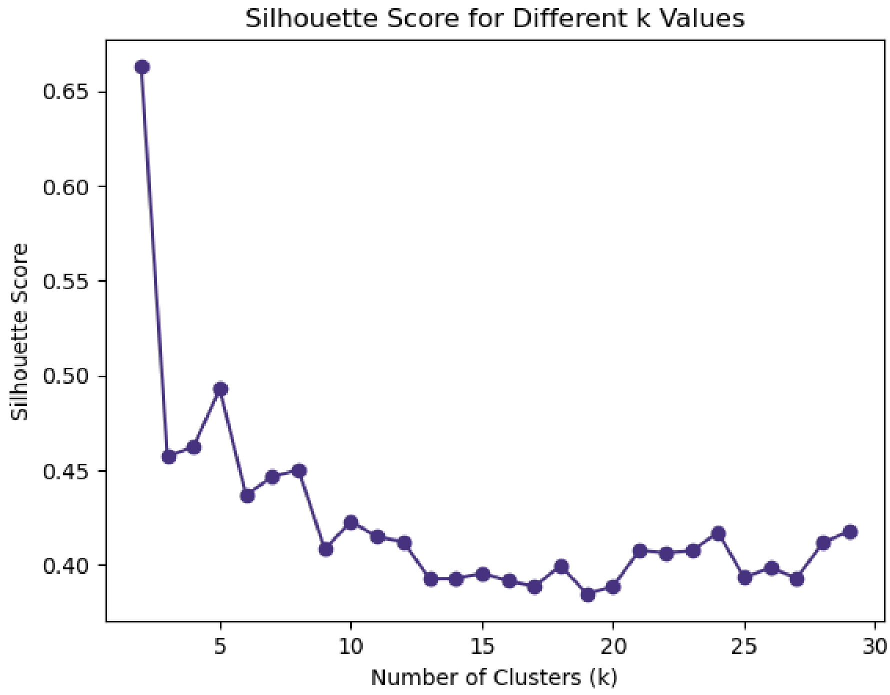
The geographical coordinates of postcodes were extracted by employing the Nominatim geolocation service from the Geopy library. Nomination initialises a geolocator to send requests to a geocoding API with a request timeout of 10 seconds to prevent a timeout. Geocoding was then iterated for all addresses in the ‘postcode’ column, storing the geocode information in a new column. The geocoder was unable to find addresses for 6 of the given postcodes, so these were dropped from the data frame. Latitude and longitude values were then extracted from the ‘geocode’ column using list comprehensions, and two new columns ‘latitude’ and ‘longitude’ were added to the data frame. The redundant geocode column was then dropped. The new data frame contained 9600 observations with 42 columns.

In order to initialise a spatial weights matrix using K-Nearest Neighbours, the optimal number of clusters first had to be determined via the elbow method. The objective was to construct a spatial lag model similar to that of Xia, Z., Stewart, K. and Fan, J (2021) to optimize the baseline models. The spatial lag model captures spatial associations among data points in a given geographical context.

A range of *k* values from 2-30 was tested. The longitude and latitude from the GeoDataFrame were stored in an array, and an empty list of 'silhouette scores' was used to store the silhouette score from different values. For each specified value from 2-30, a KMeans clustering model was produced, the model was fitted to the coordinates data, and cluster labels were predicted for each point in the data. The elbow point on the plot was determined by observing the point where the rate of decrease in silhouette scores started to slow down. By analyzing the silhouette score, which reflected the quality of clusters produced and how well separated the data is, the optimal number of clusters was determined.

Figure A shows the silhouette score for different *k* values.

Figure A: Silhouette score of different k values for parameter tuning.

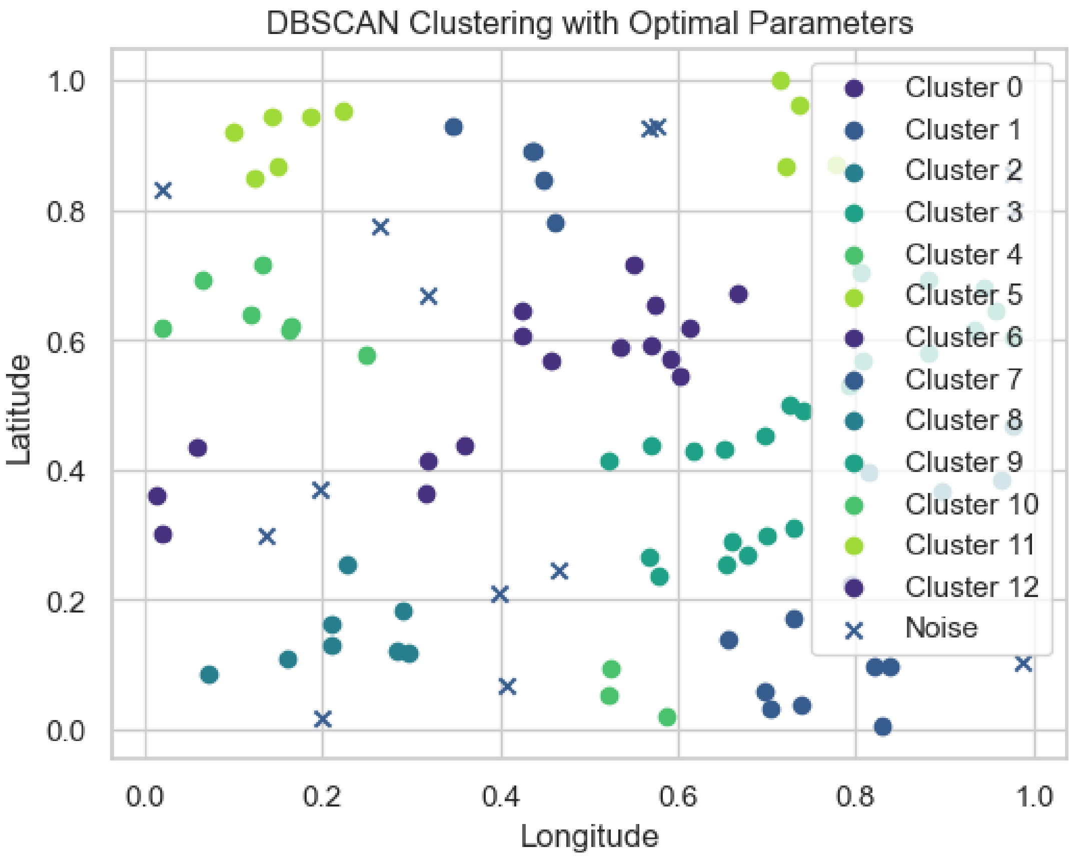


The optimal k value was determined to be 19. This choice led to a weights matrix with the fewest disconnected components, ensuring the preservation of the integrity of geospatial data. It also ensured that a substantial number of postcodes or clusters were effectively linked within the weights matrix, facilitating a comprehensive analysis of spatial relationships.

2.6 DBSCAN Clustering

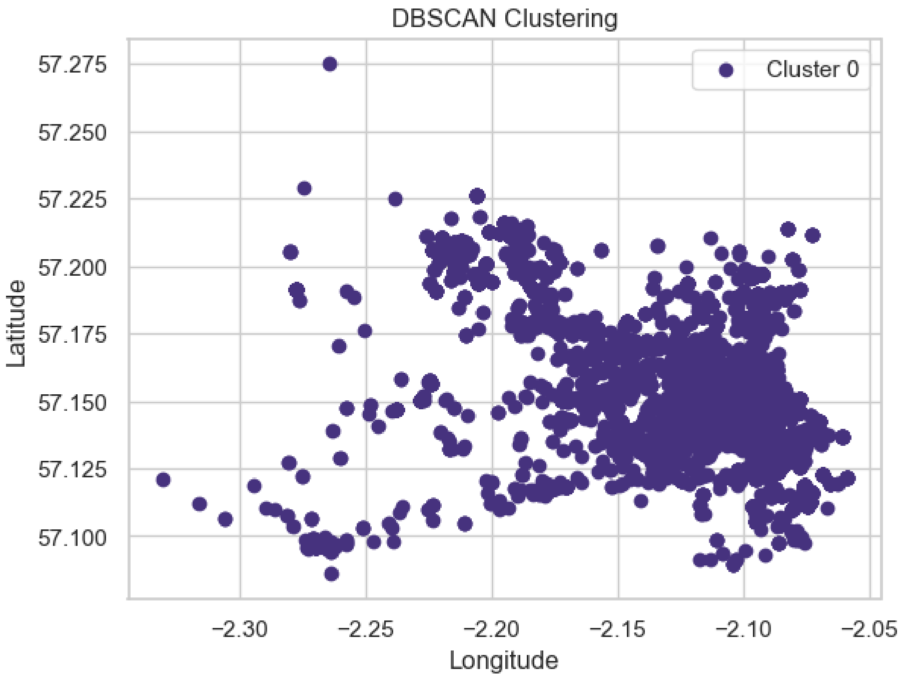
Density Spatial Based Clustering of Applications with Noise (DBSCAN) Clustering was used to further mine information from the dataset regarding geospatial data. This method was also employed by Mueller et al (2018) and Trivedi, Pardos, Sarkozy, and Heffernan (2018). Unlike traditional clustering methods that rely on assumptions about the shape of clusters, DBSCAN can identify clusters of arbitrary shapes and can also detect outliers in the data. DBSCAN creates clusters based on the density of the data points in the feature space. The algorithm defines two parameters, ‘epsilon’ and ‘min_samples. ‘min_samples‘ specifies the minimum number of points required to form a dense region or a cluster. It beings by identifying core points, which are data points within the epsilon radius and meeting or exceeding the min_samples threshold, serving as the foundation for clusters. Subsequently, the algorithm directly determines density-reachable points, where a point lies within the epsilon radius of another point, implying their inclusion within the same cluster if applicable. Clusters then emerge through chains of density-reachable points, forming the fundamental structure. The construction of clusters is initiated from core points, with density-reachable points being progressively incorporated. In cases where a point shares density-reachable status with multiple clusters, it is assigned to the cluster associated with its nearest core point. A parameter grid search was conducted to determine the optimal values for the ‘Epsilon’ and ‘MinSamples. Utilizing synthetic data, the algorithm generates combinations of parameter values, iterating through each potential cluster and the quality of clustering is evaluated using the silhouette score metric. The combination of parameters that yields the highest silhouette score is identified as the optimal choice. Figure B shows the final optimal clusters as determined by the parameter, and Figure C shows the final clusters determined.

Figure B – Optimal number of clusters for parameter tuning.



The optimal parameters for ‘epsilon’ and ‘min_samples’ were found to be 0.1 and 3.

Figure C – Final number of clusters determined.



2.7 Evaluation Metrics

In the context of evaluating classification algorithms, a set of key evaluation metrics play a pivotal role in gauging the model's performance and its ability to accurately categorize instances into different classes. These metrics offer a comprehensive framework for assessing various dimensions of the classification outcomes.

Accuracy

Accuracy quantifies the overall correctness of the model’s predictions. It measures how often the model's predictions match the actual, ground truth values in the dataset. In a classification context, where the goal is to assign input data to a specific category or class, accuracy represents the ratio of correctly predicted instances to the total number of instances in the dataset. The formula for accuracy is shown below:

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

However, it’s limited in that when there is an imbalance in class distributions, the accuracy score will not provide a complete picture of the model’s performance. As is the case with the data used in the research, other evaluation metrics are required to better evaluate the model’s performance.

Precision & Recall

Precision measures the proportion of true positive predictions made by the model, relative to the total number of positive predictions (both true positives and false positives). It is a metric that indicates how well the model minimizes false positive predictions. Recall, also known as sensitivity or true positive rate, quantifies the model's ability to identify all relevant instances of a given class. It is the ratio of true positive predictions to the total number of actual positive instances in the dataset. The formulas for calculating recall and precision are shown below:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Precision focuses on the accuracy of positive predictions among all positive predictions made by the model, while recall emphasizes the model's ability to capture all positive instances among the instances that should have been classified as positive. These two metrics often have a trade-off relationship; improving precision might lead to a decrease in recall and vice versa.

F-1

The F-1 score combines the mean of precision and recall. By balancing the trade-off of the two values, it takes both false positives and false negatives into account. The

F1 score is a industry standard metric for evaluating the performance of a model. The formula for the F1 score is shown below:

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

The F1 score, as a singular metric, effectively encapsulates the equilibrium between precision and recall. This makes it particularly advantageous in scenarios where maintaining a balance between minimizing false positives and false negatives holds paramount importance.

2.8 Test, Train, Validation Split

Upon analysing the dataset, an 80/10/10 split was decided for test/train/validation. This split ensures an ample selection of data for training, evaluation, and hyperparameter tuning. The validation set allows for hyperparameter tuning of ADASYN, geospatial weights, and clustering, and the test set serves as an unbiased benchmark that allows to measure the model’s true performance. This should deliver a comprehensive assessment of all models.

3 Analysis

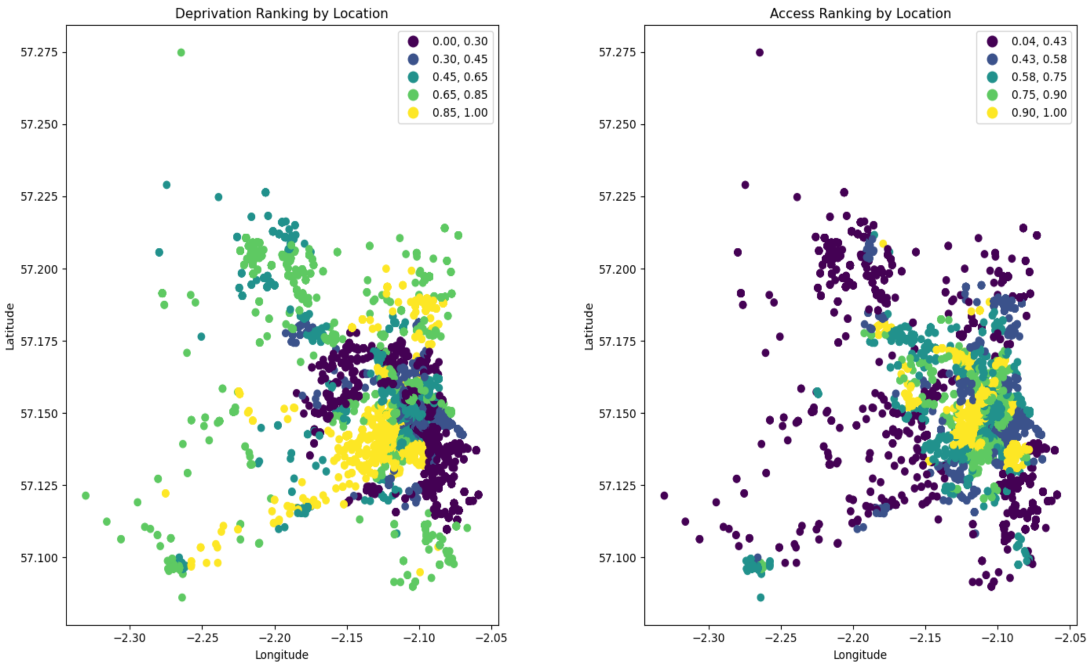
This section will outline the exploratory data analysis for the final data frame, the performance of the baseline models, the enhanced baseline models after model optimisation, and the final ensemble model using geospatial data, regression results and the optimal parameters found in the baseline model.

3.1 Exploratory Spatial Data Analysis

3.2 Deprivation Vs Access

Figure D shows the deprivation and access ranking by postcode with the values recorded in quantiles. 0.00 is the most deprived, and 1.00 is the least deprived. As all postcodes with no associated business were dropped from the final data frame, this also represents businesses by deprivation ranking and access ranking.

Figure D: Deprivation and Access Ranking by Postcode in quantiles.



The area with the most deprived clusters in Aberdeen rests between -2.15 and -2.10 longitude, and 57.125 and 57.150 latitude, with the largest cluster resting in the middle of least deprived clusters. In terms of access ranking, the most deprived clusters lie outside the main cluster, tending to cover most of the rural areas (See Appendix B). It’s worth noting that areas experiencing the highest degrees of deprivation do not invariably correspond with those facing acute accessibility challenges. Postcodes with similar deprivation rankings tend to cluster together. Postcodes with similar access rankings also tend to cluster together, but there are some areas in which more tight clusters are one or two quintiles separate. This correlates with certain business activities (See Figure *). This indicates that certain business types tend to cluster together, both in urban and rural areas. There is also evidence that some businesses are in postcodes that are not connected at all.

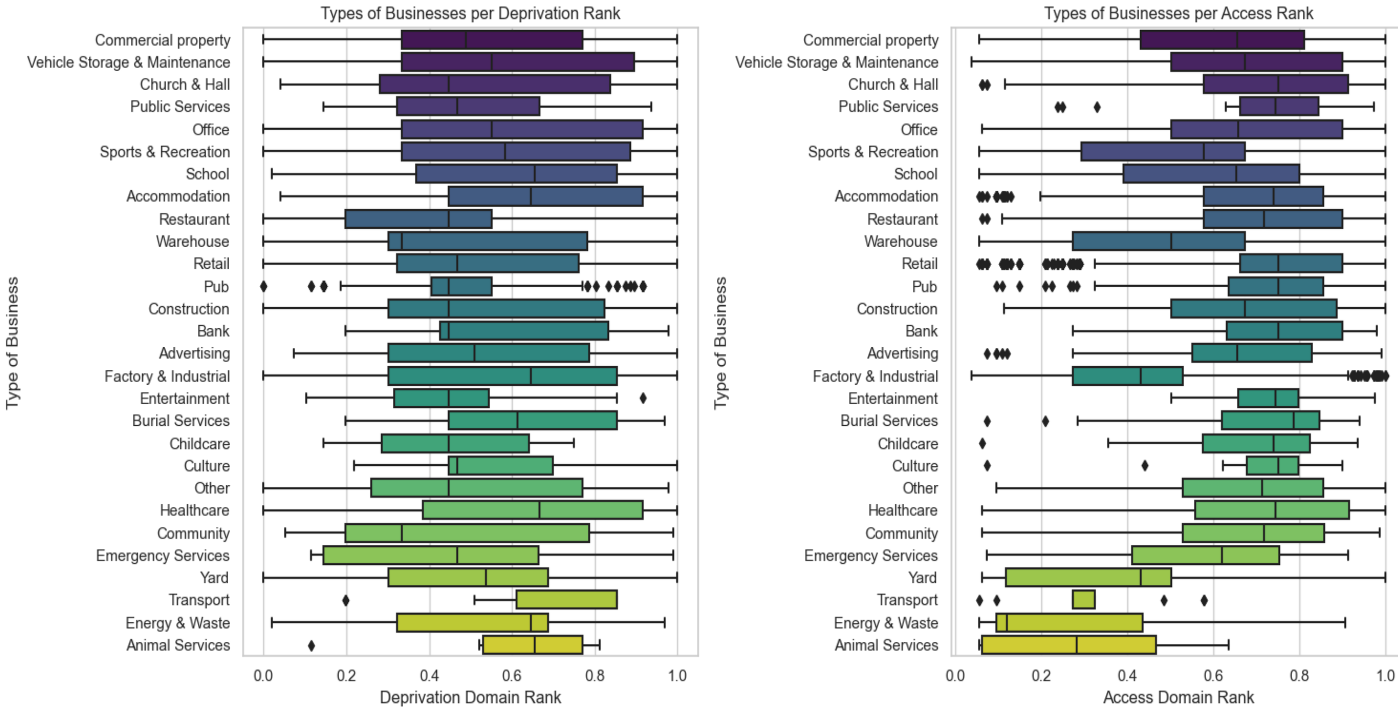
The observation that postcodes falling within the third, fourth, or fifth quintiles for deprivation tend to align with the fourth or fifth quintiles for access could potentially be attributed to the weighting methodology employed by the SIMD. This dynamic might imply that regions facing diminished access are not accurately categorized as deprived, a point that echoes criticisms of the SIMD's approach (McCartney & Hoggett, 2023) and Clelland et al (2019). This disparity could, in turn, have a cascading effect on the implementation of policies aimed at reducing deprivation within these specific areas. Despite the acknowledged lack of access, it is observed that a substantial number of businesses continue to operate outside the primary cluster, particularly in rural areas. Therefore, it is entirely possible to assess the needs of the population surrounding that area, if any, and influence policy-making decisions in order to create an educational or skills program that would cater to these needs of the business, thus simultaneously raising access, skills and education, and income deprivation. This influence could facilitate the creation of educational or skills-based initiatives specific to the requirements of local businesses, thereby simultaneously addressing issues of access, education and skills, and income-related deprivation. The anticipated outcome of such efforts would ideally create an increased willingness among businesses to establish a presence within the region, thus creating more local businesses to stimulate the economy.

A total of 22 postcodes were identified as outliers within the dataset. The pattern of clusters indicates the presence of regions characterized by sparsely connected postcodes. In certain cases, some businesses have no connection to each other at all, thus failing to form any connections in the weights matrix. Specifically, they were unable to connect to the weights matrix due to a lack of neighbouring associations up to the 19th degree, which aligns with the k value established through the K-Means parameter selection. To prevent these non-connected postcodes from exerting undue influence on the results, a proactive measure was taken to identify and subsequently remove these 22 postcodes from the data frame.

3.3 Distribution of Types of Businesses

Figure E shows the distribution of the types of businesses per deprivation and access ranking.

Figure E – Distribution of types of businesses per deprivation and access ranking.



The left plot shows the distribution of the type of businesses for deprivation ranking, whereas the right plot shows the distribution of types of businesses per access ranking. Both plots demonstrate insights into how different types of businesses are distributed among different access and deprivation rankings. The boxes show the interquartile range (IQR) of the numbers of businesses, with whiskers to show potential outliers.

There are very few businesses overall in the most deprived quartiles (0.0-0.2), which suggests a correlation between the economic conditions of an area and the presence of businesses, reinforcing Jayawarna et al (2011). Similarly, the quartile with the second least number of businesses appears to be the least deprived (0.8-1.0) although Commercial property, Vehicle Storage & Maintenance, Office, Sports & Recreation, Schools, Accommodation and Healthcare tend to gravitate towards the least deprived areas. This indicates that areas with the least deprivation provide favourable conditions for business establishment and growth, utilise high-level skills education and likely produce more income. The middle quartiles (0.2-0.8) with the middle quartiles (0.2-0.8) having a well-rounded distribution of all businesses and services.

Access ranking has an interesting effect on the type of businesses in the area. The most access-deprived quartile (0.0-0.2) has the most businesses and services belonging to the Yard, Energy & Waste, Animal Services, Factory & Industrial, and Warehouse sectors, whilst virtually none of these sectors are present in the most deprived areas. This potentially highlights an issue of clustering of businesses that may not require much manpower or skill to maintain, thus lowering the income ranking for the area. Aiming policy intervention at increasing the number and type of businesses for the most access deprived areas, including Transport, which is completely non-existent in the lowest two quartiles, will improve economic activity in the area. Additionally, a lack of such fundamental businesses and services in the least deprived areas may lead to a clustering of particular businesses (See Appendix B). However, this overall lack of businesses will have economic impacts on the area. Access and deprivation rank appear to have little correlation to the type of business found in the median quartiles (0.2-0.8). In the least access-deprived quartile (0.8-1.0), the type of businesses reflect those found in the least deprived quartiles.

The distribution of businesses exhibits an imbalance both in terms of accessibility and deprivation quartiles. Notably, the access domain reveals a higher number of outliers, which underscores an evident disparity along with the presence of smaller yet frequent clusters in certain regions. This phenomenon might arise due to the prevalence of similar businesses and services in rural areas (See Appendix C). As a result, these areas tend to generate comparable clusters on the map, indicating a recognizable geographic pattern. For instance, postcodes situated within larger clusters often encompass a range of amenities such as churches, public services, accommodations, restaurants, shops, and pubs. This alignment with the typical spatial arrangement of settlements in Western culture is noted (Phillipson, J *et al.* (2019). However, it's important to note that the current classifiers utilized in this research lack the capability to effectively address this type of scenario. As a response, the ADASYN technique was employed to counteract this issue.

4 Results

4.1 Baseline Model Results

Table 1 shows the results of the Random Forest Classifier, Gradient Boosting, Decision Tree, and Support Vector Machine baseline models.

Table 1: Baseline Model Results

	Accuracy	**Precision**	**Recall**	**F1**
Random Forest	0.41	0.32	0.41	0.35
Decision Tree	0.41	0.32	0.41	0.35
Gradient Boost	0.42	0.33	0.42	0.35
Support Vector Machine	0.41	0.29	0.41	0.33

The initial performance of the baseline models appears to be moderate due to the absence of hyperparameter tuning, which was withheld given the complexity of the research question. The Random Forest Classifier (RFC) achieved an accuracy of 41% in correctly predicting cases, with precision, accuracy, and recall scores ranging around 0.32-0.35, indicating a moderate level of performance. This suggests that the models struggled to capture intricate relationships within the dataset, potentially due to its complexity. Both the Gradient Boosting and Decision Tree classifiers exhibited similar and consistent performance levels, maintaining a moderate outcome. The Support Vector Machine (SVM) demonstrated comparable accuracy to the other models but displayed a slight drop in precision by 0.03, implying potential limitations in handling the dataset's complexity. The baseline models' performance aligns with findings from a previous study by Arribas-Bel et al (2017), where Random Forest and Gradient Boosting models showed similar results, with a Random Forest accuracy of 0.39 and Gradient Boosting at 0.29, indicating that the data has been preprocessed correctly and the models are working as expected.

4.2 Baseline Model & Spatial Weights Matrix

The weights matrix implemented using the coordinates and KNN nearest neighbours was then applied to the baseline models. Table 5 shows the results of the baseline model and the applied spatial weights matrix.

Table 5: Results of Baseline Models & *Spatial Weights Matrix*

	Accuracy	**Precision**	**Recall**	**F1**
Random Forest	0.44	0.39	0.44	0.40
Decision Tree	0.44	0.40	0.44	0.40
Gradient Boost	0.43	0.39	0.43	0.38
Support Vector Machine	0.39	0.57	0.39	0.30

RFC, Decision Tree, and Gradient Boosting models exhibited consistency with the baseline results, implying that the inclusion of spatial information through the matrix did not significantly alter their performance. Their accuracy, precision, recall, and F1 scores remained in line with baseline values, indicating limited improvements from the spatial weights. In contrast, the Support Vector Machine (SVM) model demonstrated a mixed response. While the SVM exhibited increased recall, indicating better identification of the minority class with spatial patterns, its precision decreased, implying that spatial weights led to both improved and incorrect classification of instances, predominantly in the minority class. Consequently, the SVM's overall accuracy decreased. This highlights the intricate relationship between spatial information and model performance, showcasing distinct effects across models and classes.

4.3 Baseline Model & ADASYN Results

ADASYN was utilized to balance the highly imbalanced columns related to the types of businesses and the urban-rural classification. The goal was to mitigate potential underfitting effects, with the hypothesis that the imbalance could impact model performance. Table 6 shows the results of the above baseline model and ADASYN.

Table 6: Results of Baseline Models & ADASYN

	Accuracy	Precision	Recall	F1
Random Forest	0.71	0.72	0.71	0.7
Decision Tree	0.67	0.68	0.67	066
Gradient Boost	0.61	0.61	0.61	0.59
Support Vector Machine	0.39	0.36	0.39	0.34

The outcomes of applying ADASYN were varied across the baseline models. The Random Forest Classifier (RFC) demonstrated improved accuracy of 0.53, along with enhanced precision, recall, and F1 score. ADASYN appeared to assist RFC in making more accurate predictions for minority classes. Similarly, the Decision Tree Classifier (DTC) exhibited improved metrics similar to RFC, indicating that the increase in minority samples facilitated the creation of more informative subsets for learning. Gradient Boosting Classifier (GBC) performance metrics remained relatively consistent with the baseline, suggesting that ADASYN had a limited effect on improving the model's performance in this case. However, the SVM's response to ADASYN was mixed, with a slight reduction in accuracy and F1 score, but notable improvements in precision and recall. This might be due to the influence of synthetic samples on the optimal margins that SVM aims to maximize, introducing complexity that could impact the decision boundaries of the SVM model.

Overall, the application of ADASYN resulted in more substantial improvements in model performance compared to the introduction of the spatial weights matrix. This highlights the critical role of addressing data imbalance through tailored techniques like ADASYN, which can significantly impact the predictive capabilities of machine learning models, especially for minority classes.

4.4 Baseline Model, ADASYN & Spatial Weights Matrix

Table 7 presents the integrated results of combining the baseline model with both ADASYN and the spatial weights matrix. This approach aimed to capitalize on the strengths of both techniques to achieve improved model performance.

Table 7: Results of Baseline Model + ADASYN + Spatial Weights Matrix

	Accuracy	**Precision**	**Recall**	**F1**
Random Forest	0.71	0.72	0.71	0.70
Gradient Boost	0.61	0.61	0.61	0.59
Decision Tree	0.67	0.68	0.67	0.66
Support Vector Machine	0.39	0.36	0.39	0.34

The outcomes of this integrated approach show notable improvements in model performance compared to the individual techniques. The RFC demonstrated substantial enhancement across accuracy, precision, recall, and F1 scores, reaching values of 0.71, 0.72, 0.71, and 0.70, respectively. This indicates that the combination of ADASYN and the spatial weights matrix has improved RFC's ability to predict minority classes more accurately. Similarly, the DTC achieved improved metrics across the board, with accuracy reaching 0.67 and precision, recall, and F1 scores also experiencing enhancements. Although the GBC saw a slight improvement in accuracy and precision, its recall and F1 scores remained relatively consistent. The Support Vector Machine SVM exhibited minimal improvement in recall and a slight decrease in precision, leading to marginal changes in accuracy and F1 score. Overall, the integrated approach of combining ADASYN and the spatial weights matrix demonstrated significant improvements in model performance, particularly in RFC and DTC. This underscores the potential of ensemble techniques to address complex challenges in machine learning and their effectiveness in improving the predictive capabilities of models when dealing with imbalanced and spatial data.

4.5 Baseline, Spatial Matrix, & Clustering

Table 8 shows the results of integrating the baseline model with the spatial weights matrix and clustering techniques. This comprehensive approach aimed to leverage both spatial relationships and underling data patterns to enhance the performance of the baseline machine learning models.

Table 8 – Results of baseline, spatial weights matrix, and clustering model.

	Accuracy	Precision	Recall	F1
Random Forest	0.44	0.40	0.444	0.40
Gradient Boost	0.42	0.39	0.42	0.38
Decision Tree	0.43	0.40	0.43	0.39
Support Vector Machine	0.39	0.57	0.39	0.30

The results highlight mixed performance when combining the spatial weights matrix with clustering techniques. Across all models, there is no significant improvement in accuracy, precision, recall, or F1 scores. The Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC), and Decision Tree Classifier (DTC) demonstrated only marginal changes in performance, suggesting that the combination of spatial information and clustering did not substantially enhance their predictive capabilities. Interestingly, the Support Vector Machine (SVM) showed consistent results with previous findings, exhibiting improved recall but decreased precision, indicating the potential impact of the spatial matrix on capturing spatial patterns in the minority class, but also leading to misclassification of majority classes. Overall, the integration of the spatial weights matrix and clustering techniques did not result in notable improvements in model performance across the evaluated metrics. The observed lack of substantial improvements in model performance when combining the spatial weights matrix and clustering techniques could potentially be attributed to an incompatibility between these two methods. It's possible that the clustering algorithm and the geospatial weights matrix might be interacting in a way that hinders their effectiveness rather than enhancing it.

4.6 Ensemble Baseline with Soft Voting & Spatial Weights & Cluster

The integration of ensemble techniques with the baseline model, spatial weights matrix, and clustering methods is presented in Table 9, aiming to test the relationship between the spatial weights and clustering and determine if an ensemble model would result in improved performance.

Table 9: Results of Ensemble Baseline, Spatial Weights, & Cluster

	Accuracy	**Precision**	**Recall**	**F1**
Ensemble	0.45	0.40	0.45	0.40

The results indicate that the ensemble approach did not lead to significant improvements in model performance compared to the individual components. The ensemble model's accuracy, precision, recall, and F1 score remain consistent with the baseline model results, suggesting that the combination of spatial weights, clustering, and ensemble techniques did not result in substantial enhancements in predictive capabilities.

4.7 Ensemble Baseline with Soft Voting & Spatial Weights & ADASYN

	Accuracy	**Precision**	**Recall**	**F1**
Ensemble	0.69	0.70	0.69	0.68

The results indicate that the ensemble approach led to notable improvements in model performance when combining the spatial weights matrix and the ADASYN method. The ensemble model achieved an accuracy of 0.69, along with improved precision, recall, and F1 scores, compared to the individual components. This suggests that the ensemble of spatial information and oversampling techniques effectively addressed the challenge of handling imbalanced data and capturing spatial patterns in the dataset.

The successful outcomes of this ensemble model highlight the potential benefits of combining complementary techniques to enhance predictive capabilities. By leveraging the strengths of spatial information provided by the weights matrix and the enhanced data balance achieved through ADASYN, the ensemble approach demonstrated its ability to better capture the complexities of the data and improve the model's ability to accurately predict minority classes. This underscores the importance of thoughtful technique selection when constructing ensemble models for tackling complex challenges in machine learning.

5 Conclusions

In summary, the conclusions drawn from this study reveal a nuanced picture of the interplay between various techniques applied to address the challenge of predicting business distributions within deprived regions. The integration of the spatial weights matrix into the baseline models brought about divergent outcomes. In particular, the RFC, DTC, and GBC models displayed either consistent or slightly diminished performance across key metrics like accuracy, precision, recall, and F1 scores. Notably, the SVM model demonstrated improved recall but at the cost of reduced precision, revealing a trade-off that complicates model assessment. The application of ADASYN emerged as the standout contributor to heightened model performance compared to the incorporation of the spatial weights matrix. This emphasizes the pivotal role of tailored techniques, such as ADASYN, in addressing data imbalances that substantially impact the predictive capabilities of machine learning models, particularly for underrepresented classes.

These findings underscore the intricate nature of combining multiple techniques and accentuate the demand for meticulous selection, fine-tuning, and compatibility testing when crafting ensemble models. While ensemble approaches have the potential to capture a broader spectrum of data information and relationships, the current study's results suggest that their effectiveness can be influenced by the intricacies of incorporating spatial information and clustering into the machine learning framework. Further exploration, experimentation, and methodological refinement will be crucial to unravel more potent strategies for harmoniously merging these techniques to achieve amplified model performance.

A comprehensive exploration of various machine learning techniques in the context of predicting business distributions within deprived regions shows valuable results in the interplay between the methods. The exploratory data analysis has shed light on the intricate relationship between access, deprivation, and the distribution of various types of businesses within different geographical areas, highlighting the importance of McCartney & Hogget (2023) & Clelland et al (2019), and asserting the claim that the SIMD misses adequate measurement of deprived individuals in remote and rural areas. The use of postcodes instead of data zones allowed this relationship to be explored at a more granular level and revealed previously unseen relationships. The initial baseline models provided a foundational perspective on the dataset's complexities, revealing moderate predictive capabilities hampered by the intricate relationships within the data. These findings resonated with Arribas-Bel et al. (2017) confirming the challenges of capturing spatial relationships in classification. The introduction of ADASYN proved to be a standout contributor to the mitigation of data imbalance, substantially improving model performance. This technique's efficacy underscored the significance of tailored approaches in addressing underrepresented classes, mirroring the research of Besson et al (2022) and Pradhan & Sameen (2019). The ensemble approaches showcased the potential benefits of amalgamating complementary techniques, with the combination of ADASYN and spatial weights matrix demonstrating notable improvements. The observed mismatch between clustering and geospatial techniques underscores the paramount importance of thoughtful selection and refinement of each technique to ensure their harmonious integration. It is plausible that alternative clustering algorithms or adjustments to the geospatial weight’s matrix parameters could lead to a more meaningful fusion of these techniques, thereby enhancing model performance. Delving deeper into the underlying causes of these observed outcomes and cultivating a more harmonious approach to amalgamating spatial information and clustering necessitates further investigation and experimentation.

5.1 Recommendations

Based on the findings of this study, there are several key recommendations put forward:

1) *Economic development and disparity reduction*. The insights of this study can be utilised by the Scottish Government and urban planners to target policies that foster economic growth whilst reducing disparities across the regions. By

understanding the unique characteristics of an area conducive to business establishment and identifying unique clusters of businesses within specific domains provides a basis for targeted interventions, such as skills or educational development targeted at the businesses within the region.

2) *Fine-grained deprivation assessment*. The use of postcodes instead of data zones is recommended for a more accurate and in-depth assessment of deprivation across interconnected domains. By adopting postcodes, government agencies and organizations can gain a more granular understanding of socioeconomic disparities and tailor interventions more effectively to address specific needs in different geographical areas.

3) *Comprehensive business evaluation*. The findings underscore the significance of considering both access and deprivation factors when evaluating the viability of establishing or expanding businesses. For businesses looking to invest or grow, the analysis from this study can serve as a valuable guide for assessing the potential success of ventures within diverse socioeconomic contexts. It encourages businesses to factor in spatial patterns, access, and local deprivation levels to make informed decisions about their operations.

5.2 Future Work

In addition to the comprehensive findings of this study, the results open several different paths into future research.

1) *Exploring Deep Learning Algorithms*. Future research could focus on applying deep learning algorithms to geospatial analysis. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown promise in capturing complex spatial relationships. Investigating their potential for enhancing predictive capabilities in business distribution prediction within deprived regions could lead to more accurate and robust models.

2) *Refinement of Clustering Techniques*. Given the potential pitfalls observed in the clustering techniques' application, further research could delve into reevaluating

and refining clustering methods. Exploring alternative algorithms and parameter configurations might help mitigate misinterpretation of business population clusters, especially in cases where the types of businesses vary significantly within smaller regions.

3) *Deep Dive into Clustering-Weight Interaction*. A more comprehensive examination of the interaction between clustering techniques and geospatial weights matrices is warranted. Understanding how these methods can be better aligned and fine-tuned to work harmoniously in ensemble models could enhance the predictive capabilities of such models.

4) *Optimal Cluster Size Determination*. The findings imply that the chosen static cluster size aligning with the number of base learners was suitable for this study. However, for datasets of varying sizes, it's crucial to reassess the number of clusters to ensure optimal model performance. Future research could delve into methodologies for determining the most appropriate cluster size, considering the dataset's complexity and dimensionality.

5.3 Critical Reflection

In reflecting on the dissertation, several critical considerations regarding the methods employed come to light. The study acknowledges that while predictive modelling offers valuable insights, it must be contextualized within a broader framework that addresses systemic issues and structural barriers. Therefore, it’s important to acknowledge that this study most likely did not capture the complexity and entirety of the interdependent relationships in deprivations. Relatedly, the recognition of limitations within the Scottish Index of Multiple Deprivation (SIMD) serves as a crucial reminder that the effectiveness of any model depends on the quality and reliability of its underlying data sources. The manual recategorization of businesses, while necessary, introduces subjectivity and potential errors into the classification process. The acknowledgement that certain categories like 'huts' or 'cold stores' pose challenges in classification underlines the need for more precise and standardized categorization methodologies. Finally, the spatial weights matrix's derivation from K-nearest neighbours using longitude and latitude is a methodological choice with implications. The awareness that this approach may not be optimal for spatially dispersed data suggests a potential avenue for refining the technique or exploring alternative methodologies.

In summary, the comprehensive exploration of machine learning techniques in predicting business distributions within deprived regions revealed valuable insights into the interplay between methods. The study's exploratory data analysis illuminated the complex relationship between access, deprivation, and business distribution across geographical areas, emphasizing the importance of utilizing postcodes for a finer assessment of deprivation. The initial baseline models uncovered moderate predictive capabilities limited by intricate data relationships, aligning with earlier research findings. The incorporation of ADASYN significantly improved model performance, showcasing the importance of tailored approaches for addressing underrepresented classes. Ensemble approaches demonstrated the benefits of merging techniques, with the ADASYN-spatial weights matrix combination yielding notable enhancements. However, the interaction between clustering and geospatial techniques showcased the need for meticulous selection and refinement of methods for harmonious integration. The study's recommendations include utilizing insights for economic development, adopting postcodes for deprivation assessment, and considering both access and deprivation for business evaluation. Future research avenues involve exploring deep learning algorithms, refining clustering techniques, investigating clustering-weight interactions, and determining optimal cluster sizes. Reflecting on the study, it's acknowledged that predictive modelling must consider broader societal issues, data limitations, and categorization challenges. The spatial weights matrix's derivation and manual categorization introduce potential complexities. These reflections provide a foundation for methodological refinement and future research directions.

6 References

Ali, A. et al. (2016) ‘Big data for development: Applications and techniques’, Big Data Analytics, 1(1). doi:10.1186/s41044-016-0002-4.

Arribas-Bel D, Patino JE, Duque JC (2017) Remote sensing-based measurement of Living Environment Deprivation: Improving classical approaches with machine learning. PLoS ONE 12(5): e0176684. https://doi.org/10.1371/journal.pone.0176684

Besson, M. et al. (2022) ‘Towards the fully automated monitoring of Ecological Communities’, Ecology Letters, 25(12), pp. 2753–2775. doi:10.1111/ele.14123.

Bonikowski B, DiMaggio P. 2016. Varieties of American popular nationalism. Am. Sociol. Rev. 81:949–8

Breiman, L. (2001) ‘Random Forests’, Machine Learning, 45(1), pp. 5–32. doi:10.1023/a:1010933404324.

Business rates statistics (2020) Aberdeen City Council. Available at: https://www.aberdeencity.gov.uk/services/services-business/pay-business-rates/business-rates-statistics (Accessed: 16 August 2023).

Carstairs, V. and Morris, R. (1989) ‘Deprivation: Explaining differences in mortality between Scotland and England and Wales.’, BMJ, 299(6704), pp. 886–889. doi:10.1136/bmj.299.6704.886.

Clelland, D. and Hill, C. (2019) ‘Deprivation, policy and Rurality: The limitations and applications of area-based deprivation indices in Scotland’, Local Economy: The Journal of the Local Economy Policy Unit, 34(1), pp. 33–50. doi:10.1177/0269094219827893.

Cortes, C. and Vapnik, V. (1995) ‘Support-Vector Networks’, Machine Learning, 20(3), pp. 273–297. doi:10.1007/bf00994018.

Exeter, D.J., Boyle, P.J. and Norman, P. (2011) 'Deprivation (im)mobility and cause-specific premature mortality in Scotland', Social Science &amp; Medicine, 72(3), pp. 389–397. doi:10.1016/j.socscimed.2010.10.009.

Friedman, J.H. (2001) 'Greedy function approximation: A gradient boosting machine.', The Annals of Statistics, 29(5). doi:10.1214/aos/1013203451.

Friedman, J.H. (2002) 'Stochastic gradient boosting', Computational Statistics &amp; Data Analysis, 38(4), pp. 367–378. doi:10.1016/s0167-9473(01)00065-2.

Fu, X., Feng, L. and Zhang, L. (2022) 'Data-driven estimation of TBM performance in soft soils using density-based spatial clustering and random forest', Applied Soft Computing, 120, p. 108686. doi:10.1016/j.asoc.2022.108686.

Géron, A. (2023) Hands-on machine learning with scikit-learn, keras and tensorflow: Concepts, tools, and techniques to build Intelligent Systems. 3rd edn. Sebastapol, CA: O'Reilly.

Goldberg A. 2011. Mapping shared understandings using relational class analysis: the case of the cultural omnivore reexamined. Am. J. Sociol. 116:1397–436

H. Hedegaard, B.A. Bastian, J.P. Trinidad, M.R. Spencer, M. Warner. Regional differences in the drugs most frequently involved in drug overdose deaths: United States, 2017. National Vital Statistics Reports; vol 68 , no. 12no. National Center for Health Statistics, Hyattsville, MD (2019)

Introduction to random forest in machine learning (2019) Section. Available at: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/ (Accessed: 10 August 2023).

Jayawarna, D., Jones, O. and Macpherson, A. (2011) 'New Business Creation and Regional Development: Enhancing Resource acquisition in areas of social deprivation', Entrepreneurship &amp; Regional Development, 23(9–10), pp. 735–761. doi:10.1080/08985626.2010.520337.

Kaur, H., Pannu, H.S. and Malhi, A.K. (2019) ‘A systematic review on imbalanced data challenges in machine learning’, ACM Computing Surveys, 52(4), pp. 1–36. doi:10.1145/3343440.

Macintyre, S., Macdonald, L. and Ellaway, A. (2008) ‘Do poorer people have poorer access to local resources and facilities? the distribution of local resources by area deprivation in Glasgow, Scotland’, Social Science &amp; Medicine, 67(6), pp. 900–914. doi:10.1016/j.socscimed.2008.05.029.

McCartney, G. and Hoggett, R. (2023) ‘How well does the Scottish index of multiple deprivation identify income and employment deprived individuals across the urban-rural spectrum and between local authorities?’, Public Health, 217, pp. 26–32. doi:10.1016/j.puhe.2023.01.009.

Mueller, E. et al. (2018) ‘A cluster-based machine learning ensemble approach for Geospatial Data: Estimation of health insurance status in Missouri’, ISPRS International Journal of Geo-Information, 8(1), p. 13. doi:10.3390/ijgi8010013.

Paterson, L., Blackburn, L.H. and Weedon, E. (2019) ‘The use of the Scottish index of multiple deprivation as an indicator to evaluate the impact of policy on widening access to Higher Education’, Scottish Affairs, 28(4), pp. 414–433. doi:10.3366/scot.2019.0296.

Phillipson, J. et al. (2019) ‘Shining a spotlight on small rural businesses: How does their performance compare with urban?’, Journal of Rural Studies, 68, pp. 230–239. doi:10.1016/j.jrurstud.2018.09.017.

Popham, F. and Boyle, P.J. (2011) ‘Is there a “Scottish effect” for mortality? prospective observational study of census linkage studies’, Journal of Public Health, 33(3), pp. 453–458. doi:10.1093/pubmed/fdr023.

Pradhan, B. and Ibrahim Sameen, M. (2019) ‘Predicting injury severity of road traffic accidents using a hybrid extreme gradient boosting and deep neural network approach', Advances in Science, Technology &amp; Innovation, pp. 119–127. doi:10.1007/978-3-030-10374-3_10.

Quinlan, J.R. (1986) 'Induction of Decision Trees', Machine Learning, 1(1), pp. 81–106. doi:10.1007/bf00116251.

S., P. and Uthra, A. (2022) 'Adaptive synthetic oversampling algorithm for handling class imbalance in multi-class data stream classification', Journal of Computer Science, 18(7), pp. 650–664. doi:10.3844/jcssp.2022.650.664.

Sander, J. et al. (1998) Data Mining and Knowledge Discovery, 2(2), pp. 169–194. doi:10.1023/a:1009745219419.

Scottish Government (2016) Scotland's international GDP per capita ranking March 2016. Available at: https://www.gov.scot/binaries/content/documents/govscot/publications/transparency-data/2016/03/scotlands-international-gdp-ranking-2014/documents/scotlands-international-gdp-per-capita-ranking-2014-pdf/scotlands-international-gdp-per-capita-ranking-2014-pdf/govscot%3Adocument/Scotland%2527s%2Binternational%2BGDP%2Bper%2Bcapita%2Branking%2B-%2B2014.pdf (Accessed: 05 August 2023).

Scottish Government (2016) SN 6870 - Scottish Index of Multiple Deprivation Technical Report. Available at: http://doc.ukdataservice.ac.uk/doc/6870/mrdoc/pdf/6870technical_report_2006.pdf (Accessed: 05 August 2023).

Sklearn Decision Tree Classifier (2022) scikit. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html (Accessed: 10 August 2023).

Sklearn Random Forest Classification (2022) scikit. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (Accessed: 10 August 2023).

Sklearn Support Vector Machine Classification. (2022) scikit. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC (Accessed: 10 August 2023).

Sklearn.impute.KNNImputer (2018) scikit. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html (Accessed: 06 August 2023).

The Scottish Government (2022) Scottish Government Urban Rural Classification 2020, Scottish Government. Available at: https://www.gov.scot/publications/scottish-government-urban-rural-classification-2020/pages/5/ (Accessed: 16 August 2023).

The Scottish Government (2022) Scottish Government Urban Rural Classification 2020, Scottish Government. Available at: https://www.gov.scot/publications/scottish-government-urban-rural-classification-2020/pages/5/ (Accessed: 06 August 2023).

Vera Carstairs (1995). Supplement 2: Use of Deprivation Indices in Small Area Studies of Environment and Health || Deprivation Indices: Their Interpretation and Use in Relation to Health. Journal of Epidemiology and Community Health (1979-), 49(), S3–S8. doi:10.2307/25568166

Xia, Z., Stewart, K. and Fan, J. (2021) ‘Incorporating space and time into random forest models for analysing geospatial patterns of drug-related crime incidents in a major U.S. metropolitan area’, Computers, Environment and Urban Systems, 87, p. 101599. doi:10.1016/j.compenvurbsys.2021.101599.

7 Appendix

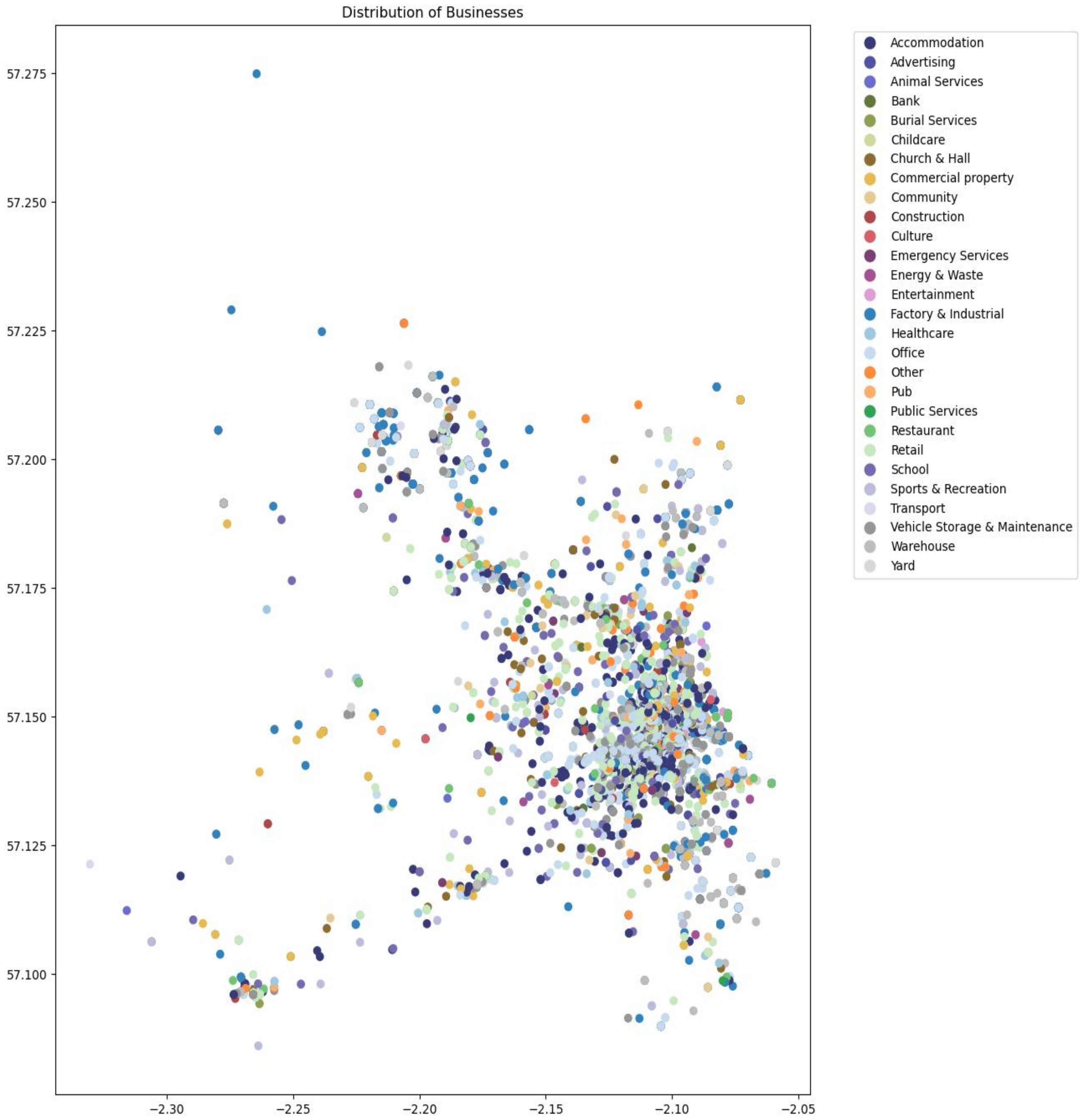
7.1 Appendix A

7.2 Appendix B

Column Name	**Description**
postcode	postcode associated with data
percentile	SIMD 2020v2 1% band - 1 is most deprived, 100 is least deprived
income_domain_rank	SIMD 2020v2 income domain rank
employment_domain_rank	SIMD 2020 employment domain rank
education_domain_rank	SIMD 2020 education domain rank
health_domain_rank	SIMD 2020 health domain rank
access_domain_rank	SIMD 2020 access domain rank
crime_domain_rank	SIMD 2020 crime domain rank
housing_domain_rank	SIMD 2020 housing domain rank
total_population	2017 NRS small area population estimates
working_age_population	2017 NRS small area population estimates and state pension age
urname	Urban Rural classification Name
income_count	Number of people who are income deprived
employment_count	Number of people who are employment deprived
cif	Comparative Illness Factor: standardised ratio
alcohol	Hospital stays related to alcohol use: standardised ratio
drug	Hospital stays related to drug use: standardised ratio
smr	Standardised mortality ratio
depress	Proportion of population being prescribed drugs for

Column Name	**Description**
PCOUT	Outward postcode
PCIN	Inward postcode
DESCRIP	Type of Business

7.3 Appendix B


Distribution of Businesses
Accommodation
Advertising
Animal Services
Bank
Burial Services
Childcare
Church & Hall
Commercial property
Community
Construction
Culture
Emergency Services
Energy & Waste
Entertainment
Factory & Industrial
Healthcare
Office
Other
Pub
Public Services
Restaurant
Retail
School
Sports & Recreation
Transport
Vehicle Storage & Maintenance
Warehouse
Yard