Big Data Research 30 (2022) 100355

Contents lists available at ScienceDirect

Big Data Research

journal homepage: www.elsevier.com/locate/bdr

Linked Open Government Data to Predict and Explain House Prices: The Case of Scottish Statistics Portal

Areti Karamanou *, Evangelos Kalampokis, Konstantinos Tarabanis

Information Systems Lab, University of Macedonia, Egnatia 156, Thessaloniki, 54636, Greece

ARTICLE INFO

Article history:
Received 31 March 2022
Received in revised form 25 July 2022
Accepted 10 October 2022
Available online 14 October 2022

Keywords:
House prices
Prediction
Gradient boosting
Explainable Artificial Intelligence

ABSTRACT

Accurately estimating the prices of houses is important for various stakeholders including house owners, real estate agencies, government agencies, and policy-makers. Towards this end, traditional statistics and, only recently, advanced machine learning and artificial intelligence models are used. Open Government Data (OGD) have a huge potential especially when combined with AI technologies. OGD are often published as linked data to facilitate data integration and re-usability. EXplainable Artificial Intelligence (XAI) can be used by stakeholders to understand the decisions of a predictive model. This work creates a model that predicts house prices by applying machine learning on linked OGD. We present a case study that uses XGBoost, a powerful machine learning algorithm, and linked OGD from the official Scottish data portal to predict the probability the mean prices of houses in the various data zones of Scotland to be higher than the average price in Scotland. XAI is also used to globally and locally explain the decisions of the model. The created model has Receiver Operating Characteristic (ROC) AUC score 0.923 and Precision Recall Curve (PRC) AUC score 0.891. According to XAI, the variable that mostly affects the decisions of the model is Comparative Illness Factor, an indicator of health conditions. However, local explainability shows that the decisions made in some data zones may be mostly affected by other variables such as the percent of detached dwellings and employment deprived population.

1. Introduction

Accurate estimation of the house prices in the housing market is valuable to various stakeholders including house owners, property investors, real estate agencies, property value assessors, lending institutions, government agencies (e.g., public housing agencies), and policy-makers. House prices may be influenced by many factors that, apart from general supply and demand conditions, include the structural or physical characteristics of the dwellings (e.g., the size of the dwelling, the number of rooms) [1,2], the marketing strategies used for the sale [3], environmental factors (e.g., air quality, safety of the neighborhood) [4], and urban amenities (e.g., parks) [5,1,6].

Literature usually employs traditional statistics to estimate the price of different types of dwellings (e.g., [7–10]). Traditional approaches, however, lack of capacity for massive data analysis, causing low utilization of data [11]. Only recently have advanced machine learning and artificial intelligence models been employed to accurately predict the price at which dwellings will be sold (e.g., [5,12–14,1,15]). These works use, apart from data that describe the structural and physical characteristics of houses, data from various other sources, such as, for example, satellite images [1,15] and point of interest (POI) data and other location information [5] from Google Maps, and social media data [14].

Open Government Data (OGD) have a huge potential especially when combined with AI technologies to bring new and fruitful insights [16]. A large part of OGD are of statistical nature and describe various indicators such as economic indicators (e.g., labor productivity), environmental indicators (e.g., greenhouse gas emissions), social indicators (e.g., employment deprivation), and health indicators (e.g., mortality ratio). OGD are often offered by various official data portals (e.g., by the official Scottish data portal[1]) as linked data. Linked data technologies facilitate the interlinking and re-usability of data coming from various sources [17] and, hence, may significantly contribute towards creating powerful predictive models.

In order, however, to be able to justify the reliability of a predictive model, it is more important to understand the factors that influence the decisions (e.g., the house prices) made by the model than achieving a good model performance [18]. Recently, eXplain-

* Corresponding author.
E-mail address: akarm@uom.edu.gr (A. Karamanou).

[1] https://statistics.gov.scot.

https://doi.org/10.1016/j.bdr.2022.100355
2214-5796/

able Artificial Intelligence (XAI) has been introduced as a set of promising techniques that could be used to understand and, hence, appropriately trust the decisions made by predictive models [19].

The aim of this paper is to create a model that predicts house prices by applying machine learning and XAI on linked OGD in order to help house owners, buyers, and investors understand which factors affect and determine the prices of houses. Towards this end, we present a case study that uses XGBoost [20], a powerful machine learning algorithm, and linked OGD from the official Scottish data portal in order to create a model that predicts the probability the mean house prices of different types of houses in the various data zones of Scotland to be higher than the average price in Scotland. The predictive model is created using OGD that can be classified in seven categories; (i) access to services, (ii) crime and justice, (iii) economic activity, benefits and tax credits, (iv) education, skills and training, (v) geography, (vi) health and social care, and (vii) housing.

The rest of this paper is organized as follows. Section 2 presents the background required to understand the case of this paper describing linked OGD and the Scottish data portal, Extreme Gradient Boosting, and XAI. Section 3 then describes the specific steps followed in this research. In addition, Sections 4 - 7 present the case study by presenting the results of OGD collection (Section 4), the exploration of data (Section 5), the creation of the predictive model (Section 6), and the explanation of the predictive model (Section 7). Section 8 discusses the results of this paper and, finally, Section 9 concludes this paper.

2. Background

This section presents the background knowledge required to understand the content of this paper. Specifically, it describes (i) Open Government Data and Linked Open Government Data including a presentation of the Scottish data portal, (ii) Extreme Gradient Boosting, and (iii) Explainable Artificial Intelligence.

2.1. Open Government Data

Open Government Data (OGD) are data published by the public sector in open and reusable formats without restriction or charge for their use by society [21]. OGD is a political priority the last decade in many countries in order to harness multifaceted benefits including enhancing evidence-based policy making and stimulating economic growth. As a result, a large number of public authorities and National Statistical Institutes internationally have already been publishing their data in their official data portals (e.g., https://www.data.gov in the U.S. and https://dati.gov.it/ in Italy). However, the current scarce use of OGD has impeded reaching the full potential of OGD [22]. The recent Europe's recast Directive (EU) 2019/1024[2] on Open Data and the Re-Use of Public Sector Information shifts the focus from just publishing OGD to sharing and re-using them.

A large part of OGD on the Web is of statistical nature, meaning that it consists of highly structured numeric data [23]. They also usually concern aggregated data that monitor demographical, social and business indicators across countries. Statistical data are multidimensional, meaning that a measure is described based on multiple of dimensions. As a result, statistical data are commonly conceptualized using the data cube model that has already been introduced for the needs of the Online Analytical Processing (OLAP) and data warehouse systems. According to literature, a data cube comprises [24–27]: (1) (one or more) measures, which represent numerical values, and (2) dimensions, which provide contextual information for the measures. For example, a dataset that measures unemployment rate in different administrative regions of European countries in years 2019 - 2021 has one a single measure, i.e., unemployment, and two dimensions, the geospatial dimension, i.e., "country", and the temporal dimension, i.e., "year". Each dimension comprises a set of distinct values. In the previous example, the "country" dimension has values "GR", "FR", etc., while the "year" dimension values 2019, 2020, and 2021. An additional dimension could be "age group" with values 00–24, 25–49, and 50+. Dimension values can be hierarchically organized into levels representing different granularity. In the example, the "country" dimension has a single hierarchical level, however there could be more, e.g., the geospatial dimension may have both countries and regions. Finally, the location of each cube's cell is specified by the dimension values, while the value of a cell specifies the measure (e.g., the unemployment rate of "GR" in "2021" is "23.1"%").

2.1.1. Linked Open Government Data

Linked data has been introduced as a promising technological paradigm for opening up data because it facilitates the integration of data. In the case of statistical data, linked data has the potential to realize the vision of performing data analytics on top of integrated but previously isolated statistical data coming from various sources across the Web [28,29].

Linked data are based on the Semantic Web philosophy and technologies and are mainly about publishing structured data using the W3C Resource Description Framework (RDF) standard. Linked data use HTTP URIs to name entities and concepts rather than focusing on the ontological level or inferencing. As a result, users can further look up entities and concepts to obtain more information. The names used follow the *prefix:localname* notation, where the prefix identifies a namespace URI. For example, *foaf:Person* is the name of the Person class of the FOAF[3] vocabulary with URI http://xmlns.com/foaf/0.1/Person.

The QB vocabulary [30] is a W3C standard that uses the linked data principles to publish statistical data on the Web. The core class of the vocabulary is the *qb:DataSet* that represents a data cube which includes a collection of observations (*qb:Observation*) (the cells of the data cube). A data cube comprises a set of dimensions (*qb:DimensionProperty*) that define what the observations apply to (e.g., gender, reference area, time, age), measures (*qb:MeasureProperty*) that represent the phenomenon which is being observed, and attributes (*qb:AttributeProperty*) that are used to represent structural metadata such as the unit of measurement.

The values of the dimensions are commonly populated using predefined and possibly hierarchical code lists. For example, a geospatial dimension can be populated by a code list that includes URIs for all geographical or administrative divisions of a country. The values of code lists are usually specified using either the QB vocabulary or the W3C standard Simple Knowledge Organization System (SKOS) vocabulary [31]. Code lists may also include hierarchical relations expressed using the SKOS vocabulary (e.g., using the *skos:narrower* property), the QB vocabulary (e.g., using the *qb:parentChildProperty*) or the XKOS[4] vocabulary (e.g., using the *xkos:isPartOf* property).

Common concepts like dimensions, measures, attributes, and code lists can also be re-used to facilitate their discovery. Towards these directions, the UK Government Linked Data Working Group has defined a set of common concepts based on the SDMX guidelines.[5] These pre-defined concepts are currently widely used al-

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2019.172.01.0056.01.ENG.

[3] http://xmlns.com/foaf/spec/.

[4] https://rdf-vocabulary.ddialliance.org/xkos.html.

[5] https://github.com/UKGovLD/publishing-statistical-data.

Table 1
Number of datasets per theme in the Scottish data portal.

Theme	No of datasets
Access to Services	3
Business, Enterprise and Energy	22
Children and Young People	4
Community Wellbeing and Social Environment	22
Crime and Justice	21
Economic Activity, Benefits and Tax Credits	46
Economy	10
Education, Skills and Training	16
Environment	14
Geography	5
Health and Social Care	61
Housing	32
Labor Force	10
Management Information	5
Population	20
Reference	2
Scottish Index of Multiple Deprivation	11
Transport	8
Total	314

though they are not part of the QB vocabulary. Example dimension concepts include *sdmx:timePeriod*, *sdmx:refArea*, and *sdmx:sex*, and measure concepts *sdmx:obsValue*.

Today, governments provide OGD on the Web through their official data portals. Examples include the Scottish data portal, the data portal of the Japan's government (e-Stat), the data portal of the environmental department of the Flemish government (VLO), DCLG in the UK, and the data portals that host the Italian (ISTAT) and Irish (Irish CSO) 2011 censuses. All the official data portals provide data as linked data through their SPARQL endpoints and thus the data can be easily collected by specifying and submitting relevant SPARQL queries. Connecting data from the data portals would create a Knowledge Graph of qualitative and fine-grained statistical data that would facilitate data discovery and collection. Towards this end, literature has already identified [32] and addressed [17] interoperability challenges for connecting statistical data from multiple trustworthy sources.

2.1.2. The Scottish data portal

The Scottish data portal[6] provides statistical data for free reuse under the Open Government Licence (OGL). The portal currently hosts 314 datasets covering various societal and business aspects of Scotland classified into 18 themes and published by 13 organizations like the Scottish Government and SEPA. Table 1 presents the number of datasets per theme.

Datasets are provided at different levels of spatial granularity in Scotland starting from the level of postcodes to the level of council areas. Table 2 presents all levels of spatial granularity used in the data portal. For example, data zones (DZs) refer to the primary geography for the release of small areas statistics in Scotland. DZs were introduced after the 2001 census and changed in 2011 after the 2011 census. There are 6976 2011 data zones covering the whole of Scotland and designed to have roughly standard populations of 500 to 1,000 household residents. At the same time, Council Areas are the coarser geographical units in Scotland. There are currently 32 Council areas in Scotland each of which is governed by a unitary council.

The Scottish data portal allows users to navigate through its pages in order to view and retrieve data as tables, maps, and charts or download them in various formats (e.g., html, json, csv). Alternatively, data can be retrieved as linked data using advanced, flexible queries that can be submitted to the SPARQL endpoint[7] released by the portal. Specifically, the portal utilizes linked data technologies to make data available as a unified knowledge graph. The different datasets are organized as data cubes and connected through typed links mainly using the RDF Data Cube vocabulary. As a result, users are able to search in a uniform way across all the available datasets and easily combine data from different datasets.

A dataset in the Scottish linked data portal usually includes multiple measures. For example, the *Employment deprivation* dataset[8] contains employment deprivation counts and rates. Each measure has its own URI of the form http://statistics.gov.scot/def/measure-properties/XXXXX, for example, http://statistics.gov.scot/def/measure-properties/count and is sub-property of *sdmx-measure:obsValue*. Then, the measure used in each observation is defined using the *qb:MeasureType* property. Regarding the dimensions of the data cubes, time, geography, age, and gender dimensions are commonly used. For time and geography, usually dimensions of the SDMX vocabulary are re-used; *sdmx-dimension:refPeriod* and *sdmx-dimension:refArea* respectively. Age and gender dimensions use their own properties, e.g., http://statistics.gov.scot/def/dimension/age for the age dimension, and http://statistics.gov.scot/def/dimension/gender or http://statistics.gov.scot/def/dimension/sex for the gender dimension.

Dimensions are commonly populated with values reused from code lists that hold their potential values. For example, the values of the spatial dimension, i.e., "reference period", may be calendar years, e.g., http://reference.data.gov.uk/id/year/2015, two-year intervals, e.g., http://reference.data.gov.uk/id/government-year/2017-2018, or three year intervals that begin at the start of the 1st minute of 1st hour of the 1st January of a Gregorian calendar year, e.g., http://reference.data.gov.uk/id/gregorian-interval/2012-04-01T00:00:00/P3Y that refers to the three year interval that starts from the 1st second of 1st minute of 1st hour of the 1st January 2012 in the Gregorian Calendar. The type of reference period is defined using the http://statistics.gov.scot/def/dimension/timePeriod property. The geography dimension also reuses values from code lists that are sometimes connected with generalization/specialization relations to form geographical hierarchies.

An example of an observation of the *Births* dataset[9] is presented in Fig. 1. The dataset measures the number of births registered in a calendar year. The observation shows that, in 2019, there were 42 male births in the Halfway, Hallside and Drumsagard - 07 DZ.

2.2. Extreme Gradient Boosting

EXtreme Gradient Boosting (XGBoost) is an implementation of a generalized gradient boosting algorithm [20]. It is a highly scalable and efficient system offered as an open source package that runs more than ten times faster than existing popular solutions on a single machine and scales to billions of examples in distributed or memory-limited settings. Boosting refers to the general problem of boosting the performance of weak learning algorithms by combining all the generated hypotheses into a single hypothesis [33]. The idea of boosting was further elaborated in gradient boosting [34], where one new weak learner is added at a time and existing weak learners in the model are frozen and left unchanged. XGBoost supports both regression and classification problems.

A key to the performance of XGBoost is properly tuning its hyperparameters. XGBoost, for example, creates trees that are based on the residuals of the previous tree. New trees are created that

[6] http://statistics.gov.scot.

[7] https://statistics.gov.scot/sparql.

[8] http://statistics.gov.scot/data/scottish-index-of-multiple-deprivation--employment-indicators.

[9] http://statistics.gov.scot/data/births.

Table 2
Levels of spatial granularity in the Scottish government data portal.

Granularity Level	Description	No of areas
Postcodes	Live postcodes in Scotland	159,187
2001 data zones	Primary geography for the release of small area statistics (2001)	6,505
2001 intermediate zones	Statistical geography that sit between data zones and council areas (2001)	1,235
2011 data zones	Primary geography for the release of small area statistics (2011)	6,976
2011 Intermediate Zones	Statistical geography that sit between Data Zones and council areas (2011)	1,279
Electoral Wards	Areas served by councillors at local government level	354
Localities	Subdivision of large settlements to reflect areas which are more easily identifiable as the towns and cities of Scotland	655
Settlements	Built-up areas of 500 people or more	519
Scottish Parliamentary Constituencies	Constituencies of the Scottish Parliament at Holyrood	73
Westminster Parliamentary Constituencies	Constituencies of Westminster	59
Health Board Areas	Responsible for the delivery of frontline healthcare services in Scotland	14
Integration Authorities	Local partnerships of NHS and local council care services that are jointly responsible for the health and care needs of patients	31
Travel to Work Areas	Approximations to self-contained labor markets reflecting areas where most people both live and work	47
Council Areas	Local government responsible for the provision of a range of public services	32
National Parks	National Parks in Scotland	2

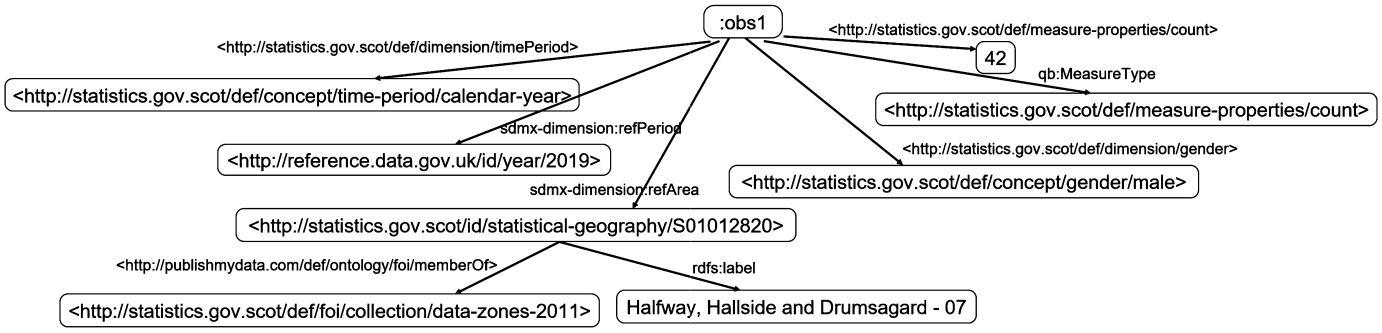


Fig. 1. An observation of the "Births" dataset of the Scottish data portal. The observation shows the number of male births (42) in 2019 in the Halfway, Hallside and Drumsagard- 07 DZ.

predict the residuals or errors of prior trees and then added together in order to make the final prediction. The final prediction is made either when the computed residuals have reached their minimum values or when the maximum number of trees is reached. As a result, XGBoost minimizes the loss in the final prediction. The number of trees boosted is indicated by the n_estimators hyperparameter and the maximum depth of each tree by the max_depth hyperparameter. Then, for the training part, it is important to properly select the subsample ratio of the training instances (subsample) which indicates the part of training data that will randomly be sampled prior to growing trees in order to prevent overfitting. XGBoost is also based on column subsampling and allows tuning, for example, the fraction of columns selected for each tree (colsample_bytree), for each depth level reached in a tree (colsample_bylevel), or for each node/split (colsample_bynode). In addition, learning_rate is used to shrink the feature weights and prevent overfitting. XGBoost also allows regularization, e.g., L1 (alpha) and L2 (lambda). Finally, scale_pos_weight is used to control the balance of positive and negative weights. This last hyperparameter is useful for imbalanced classes.

The XGBoost algorithm has been applied to many domains, such as transportation, health, and energy, because of its high speed, high accuracy, and good robustness. It is indicative that during 2015, the 17 out of 29 winning solutions that were submitted to Kaggle competitions used XGBoost [20]. In the field of house price prediction, few research works have used XGBoost (e.g., in [5,35,36]).

2.3. Explainable Artificial Intelligence

Artificial intelligence is today increasingly used to make predictions and facilitate decision making. AI produces a result (e.g., a prediction) starting from data and without providing details on the reasoning behind the results. This fact is known as "black box" [37] and hampers trust and acceptance of AI. To overcome AI "black box", eXplainable Artificial Intelligence (XAI) was introduced that aims to make the results of AI more understandable to humans and, hence, destroy the "black box" phenomenon. The term XAI was first used by Van Lent et al. [38] although the conception of the idea traces back to mid 1970s [39]. Nevertheless, XAI started to gain the interest of researchers only recently, with the penetration of AI in almost all industries. XAI has been already proved particularly important in medical applications [40–43] and in transport [44].

In general, XAI works may regard the development of predictive models that are interpretable by design, without other frameworks (e.g., [45]), or the development of model agnostic explainability frameworks that can be used on predictive models model to make them more interpretable (e.g., [46]) [47]. In the latter category fall a variety of successful XAI approaches including LIME, Anchors, XGNN, and SHAP [48]. SHapley Additive exPlanations (SHAP) is a unified framework for interpreting predictions based on Shapley values [49]. The Shapley value method is a game theory method that assigns payouts to players depending on their contribution to the total payout where players cooperate in a coalition [50]. In machine learning the "game" is the prediction task for a single instance of the dataset. The "gain" is the prediction minus the average prediction of all instances and the "players" are the feature values of the instance that collaborate to receive the gain. The Shapley value is the average marginal contribution of a feature value across all possible coalitions.

The Shapley value method is computationally expensive because going over all coalitions scales exponentially with the increase in the number of features. SHAP solved this problem by enabling the exact computation of Shapley values in low order polynomial time instead of exponential by leveraging the internal structure of tree-based models [51,52]. SHAP also proposed SHAP interaction values, which are an extension of Shapley values that directly capture pairwise interaction effects. Moreover, SHAP introduced global interpretation methods based on aggregations of Shapley values such as the SHAP feature importance, which is measured as the mean absolute Shapley values.

In terms of the scope of interpretability, SHAP has the capability to explain both overall model prediction (Global Feature Importance) and also specific prediction (Local Feature Importance). Finally, since SHAP is model agnostic, it works for several algorithms including regression, tree-based, and boosting algorithms.

3. Research approach

In order to understand the factors that affect the mean house prices we use the following steps: (1) collect data, (2) explore data, (3) create the predictive model, and (4) explain the predictive model.

(1) Collect data. In this step we collect data from the datasets of the Scottish data portal in order to create the predictive model. Towards this end, we first search through the datasets of the data portal to select a dataset appropriate to be used as the dependant variable in the predictive model. We then employ a SPARQL query to find all compatible datasets from the Scottish data portal that will be used as the predictors in the predictive model. Compatible datasets must describe indicators that match the year of reference and the spatial granularity level of the dependant variable. We, finally, use a second SPARQL query to collect the data included in the compatible datasets. We use the collected data to create the predictors of the predictive model. In datasets with multiple measures we keep only ratios.

(2) Explore data. In this step we explore data retrieved in the previous step for better understanding. Towards this end, we use descriptive statistics and visualizations of data.

(3) Create the predictive model. We use the XGBoost algorithm to create the predictive model for the classification. Collected data are randomly split to 70% training data and 30% testing data. —feature selection— The tuning of the model is performed using repeated 5-fold cross-validation on the train set. Cross-validation is selected because it ensures low variance and bias for the predictive model. Specifically, cross-validation is used on various possible values of hyperparameters in order to choose the value that gives the lowest cross-validation average error. The employed XGBoost hyperparameters are n_estimators, learning_rate, max_depth, subsample, colsample_bytree, colsample_bylevel, and scale_pos_weight. The evaluation of the model is performed on the test set after tuning the model. The performance of the predictive model is measured using (a) the sklearn classification report, (b) the Receiver Operating Characteristic (ROC) AUC score with confidence interval, and (c) the Precision Recall Curve (PRC) AUC score. The sklearn classification report calculates basic metrics like precision, recall, and f1-score. AUC measures the area underneath the entire ROC curve and PRC respectively. In general, in classification and diagnostic tests, the ROC AUC score describes how an adjustable threshold impacts false positives and false negatives. However, the ROC AUC score is only partially meaningful when used with imbalanced data [53]. We address this limitation by additionally employing the PRC AUC score to evaluate the classification model.

Moreover, we formulate the same problem as a regression problem in order to predict the 2015 house prices in Scottish 2011 data zones as real numbers. We use again XGBoost and evaluate the regression model using Root Mean Square Error (RMSE) which is calculated based on the difference between the predicted and the actual house prices.

(4) Explain the predictive model. Towards this end, we employ the SHAP framework to explain both the prediction of the model in all data zones (global explainability) and the prediction in specific data zones (local explainability). We first calculate the Shapley values for all variables using the predictive model created in the previous step. We then further fine-tune the magnitude of the predictive model to determine the optimal number of variables. We remove the least important variable based on their Shapley values one at a time to create a new model. We select the optimal predictive model with the minimal accuracy. We use this model to globally and locally explain the decisions made.

For the global explanation of the model, we use the following types of visualizations:

- SHAP summary plots that are beeswarm plots where each dot's position is determined by the feature on the y-axis and the Shapley value on the x-axis. The color represents the value of the feature from low to high.
- A bar plot with variables in order of importance based on their Shapley values. The plot also uses colors to classify variables in their theme.
- SHAP dependence plots that show how a feature's value (x-axis) impacts the prediction (y-axis) of every sample (each dot) in a dataset and how the change in SHAP values across a feature's value range.

For the local explainability we use decision plots. Decision plots present the importance of all variables of the model. The y-axis of the plot presents the variables ordered by importance based on the Shapley values while the x-axis the output of the predictive model. Each decision plot presents the importance of variables for a different Scottish data zone.

4. Collect data

In order to predict the probability the mean price of houses in each 2011 data zone in 2015 to be higher than the average mean price in all 2011 data zones (classification problem) and also the actual values of house prices in each 2011 data zone in 2015, we select data from datasets of the Scottish data portal to use for the (a) dependant variable and (b) predictor variables of the predictive model.

```
PREFIX sdmx-dim: <http://purl.org/linked-data/sdmx/2009/dimension#>
SELECT distinct ?b
WHERE {
    ?a qb:dataSet ?b;
        sdmx-dim:refArea [?m <http://statistics.gov.scot/def/foi/collection/data-zones-2011>].
        OPTIONAL{ ?a sdmx-dim:refPeriod <http://reference.data.gov.uk/id/year/2015>}.
        OPTIONAL{ ?a sdmx-dim:refPeriod <http://reference.data.gov.uk/id/government-year/2014-2015>}.
        OPTIONAL{ ?a sdmx-dim:refPeriod <http://reference.data.gov.uk/id/gregorian-interval/2014-01-01T00:00:00/P3Y>}.

}
```

Fig. 2. The SPARQL query to retrieve compatible datasets. The query searches for datasets that describe indicators about 2011 data zones and the time period 2015, 2014-2015, or 2014-2016.

Table 3
Variables.

Theme	Variables
Access to Services	Travel times (minutes) to GP surgeries by public transport, Travel times (minutes) to post office by public transport, Travel times (minutes) to retail center by public transport, Travel times (minutes) to petrol station by car, Travel times (minutes) to post office by car, Travel times (minutes) to GP surgeries by car, Travel times (minutes) to primary school by car, Travel times (minutes) to secondary school by car, Travel times (minutes) to retail center by car
Crime & Justice	Chimney fires (ratio), Dwelling fires (ratio), Other building fires (ratio), Other primary fires (ratio), Outdoor fires (ratio), Refuse fires (ratio), Vehicle fires (ratio), Accidental chimney fires (ratio), Accidental dwelling fires (ratio), Accidental other building fires (ratio), Accidental other primary fires (ratio), Accidental outdoor fires (ratio), Accidental refuse fires (ratio), Accidental vehicle fires (ratio), Not accidental chimney fires (ratio), Not accidental dwelling fires (ratio), Not accidental other building fires (ratio), Not accidental other primary fires (ratio), Not accidental outdoor fires (ratio), Not accidental refuse fires (ratio), Not accidental vehicle fires (ratio), Crime indicators (ratio)
Economic Activity, Benefits & Tax Credits	Children 0–15 living in low income families (ratio), Children 0–19 living in low income families (ratio), Age of first time mothers 19 years and under (ratio), Age of first time mothers 35 years and older (ratio), Employment deprivation (ratio)
Education, Skills & Training	School attendance (ratio), Educational attainment of school leavers (score)
Geography	Land area (in hectares), Urban Rural Classification
Health and Social Care	Mothers currently smoking (ratio), Mothers former smokers (ratio), Mothers never smoked (ratio), Low birth–weight (less than 2500 g) babies (single births) (ratio), Mothers not known if they smoke (ratio), Comparative Illness Factor
Housing	Dwellings per hectare (ratio), Detached dwellings (ratio), Flats (ratio), Semi–detached dwellings (ratio), Terraced dwellings (ratio), Dwellings of unknown type (ratio), Long–term empty households (ratio), Occupied households (ratio), Second–home households (ratio), Vacant households (ratio), Households with occupied exemptions (ratio), Households with unoccupied exemptions (ratio), Households with single adult discounts (ratio)

For the dependant variable we select the "House prices" dataset.[10] The dataset measures the *mean*, *median*, *lower quartile*, and *upper quartile* house prices in different spatial granularity levels and years. Spatial granularity levels include Countries, Council Areas, Scottish Parliamentary Constituencies, Health Board Areas, Electoral Wards, Community Health Partnerships, 2011 Intermediate Zones, 2011 Data Zones, 2001 Intermediate Zones, and 2001 Data Zones. Years of reference are all years from 1993 up to 2018. However, values of house prices are not provided for all spatial levels, years and measure types. For example, we can retrieve 2004 mean house prices only for Countries, Council Areas, Scottish Parliamentary Constituencies, 2001 Intermediate Zones, and 2001 Data Zones. We select data about mean house prices in 2011 Scottish data zones in 2015 for the dependant variable of the predictive model.

For the predictors, we select data from datasets that are compatible with the "House prices" dataset. To this end, we submit two SPARQL queries to the Scottish data portal to get relevant data as linked data; the first one to get all datasets that are compatible with the "House prices" dataset, and the second one to retrieve data from the compatible datasets. The two main requirements to identify compatible datasets are that (a) the year of reference is 2015, and (b) the granularity level of the data included is 2011 data zones. The SPARQL query for retrieving all compatible datasets is presented in Fig. 2. The query searches for datasets that use 2011 datazones as geography and years 2015, 2014-2015, or 2014-2016.

The above query results in twenty seven compatible datasets. However, the number of predictor variables that we can create from the compatible datasets is greater. For example, in the "Age of First Time Mothers" dataset[11] the "Age" dimension holds three values, i.e., (a) 19 years and under, (b) 35 years and over, and (c) all. If we lock the "Reference Period" dimension to "2014/15-2016/17" and the "Reference Area" dimension to "2011 Data Zones", then we can result in three variables, one per different value of the "Age" dimension. In addition, when datasets include more than one measures (e.g., count and ratio) we keep only ratios. We, hence, result in 59 predictor variables classified in the eight themes of the Scottish data portal (Table 3).

We then retrieve the data of the 59 selected variables. Towards this end, we submit a second SPARQL query to the Scottish data portal. Part of the query can be seen in Fig. 3. The query selects the dependent variable and two predictor variables for all 2011 data zones. Specifically, the query retrieves (a) the prices of houses in 2015 (dependent variable), (b) the ratio of employment deprivation for the years 2014-2015 (first predictor variable), and (c) the ratio of school attendance during the years 2014-2015 (second predictor variable). Null values are included in the results.

5. Explore data

In order to better understand data, we first classify the compatible datasets and the predictor variables into their theme (Table 4).

[10] http://statistics.gov.scot/data/house-sales-prices.
[11] https://statistics.gov.scot/data/age-at-first-birth.

```
SELECT ?area ?price ?employmentdeprivation ?schoolattendancerate WHERE {
{
?a qb:dataSet <http://statistics.gov.scot/data/house-sales-prices>;
<http://purl.org/linked-data/sdmx/2009/dimension#refPeriod> <http://reference.data.gov.uk/id/year/2015>;
<http://purl.org/linked-data/sdmx/2009/dimension#refArea> ?area.
?area <http://publishmydata.com/def/ontology/foi/memberOf> <http://statistics.gov.scot/def/foi/collection/data-zones-2011> .
?a <http://statistics.gov.scot/def/measure-properties/mean> ?price .
}
OPTIONAL{
  ?u qb:dataSet <http://statistics.gov.scot/data/scottish-index-of-multiple-deprivation---employment-indicators>;
  <http://purl.org/linked-data/sdmx/2009/dimension#refPeriod> <http://reference.data.gov.uk/id/government-year/2014-2015>;
  <http://purl.org/linked-data/sdmx/2009/dimension#refArea> ?area;
  <http://statistics.gov.scot/def/measure-properties/ratio> ?employmentdeprivation.
}
OPTIONAL {
    ?f qb:dataSet <http://statistics.gov.scot/data/school-attendance-rate>;
    <http://purl.org/linked-data/sdmx/2009/dimension#refPeriod> <http://reference.data.gov.uk/id/government-year/2014-2015>;
    <http://purl.org/linked-data/sdmx/2009/dimension#refArea> ?area;
    <http://statistics.gov.scot/def/measure-properties/ratio> ?schoolattendancerate.
 }
}
```

Fig. 3. Part of the SPARQL query to retrieve data from the Scottish data portal. The query retrieves data for the 2015 house prices in all data zones, along with data from two compatible datasets describing (i) the ratio of employment deprivation, and (ii) the ratio of school attendance.

Table 4
Number of selected datasets & variables per theme.

Theme	Datasets	Variables
Access to Services	1	9
Crime and Justice	2	22
Economic Activity, Benefits and Tax Credits	2	5
Education, Skills and Training	2	2
Geography	2	2
Health and Social Care	4	6
Housing	3	13
Total	16	59

Table 5
Descriptive statistics for the mean prices of houses in all 2011 data zones in 2015.

Mean	Median	Min	Max
£163,478	£142,158	£20,604	£1,244,910

Most compatible datasets regard "Health and Social Care" theme (4 datasets or 25%). Nevertheless, most predictor variables come from the "Crime and Justice" theme (22 variables or 37.2%), followed by the variables classified to the "Housing" theme (13 variables or 22%).

We then explore the data of the dependant variable. Table 5 presents the descriptive statistics for the mean prices of houses in all 2011 data zones in 2015 according to the related dataset retrieved from the Scottish data portal. The 2011 data zone with the minimum mean house price (£20,604) is Cumbernauld Central of Glasgow, while the 2011 data zone with the maximum mean house price (£1,244,910) is Leith (Albert Street) - 03, a port area in the north of the city of Edinburgh. The mean house price across all data zones in 2015 is £163,478. The problem, hence, we explore in this work is the prediction of the probability the average price of houses in each 2011 data zone in 2015 to be higher than £163,478. In 2015, 39% of the 2011 data zones (i.e., 2348 out of the 6014 data zones) have a mean price above the average of all data zones. This means that the observations between the two classes that will be predicted, i.e., data zones with mean house prices (a) above the average, and (b) below the average, are slightly imbalanced.

In addition, the map in Fig. 4 shows the mean house prices in all 2011 data zones. More expensive houses are represented in the map with the darker red highlights. Northern data zones of Scotland have less expensive houses, while the mean house price in the majority of western data zones is increased.

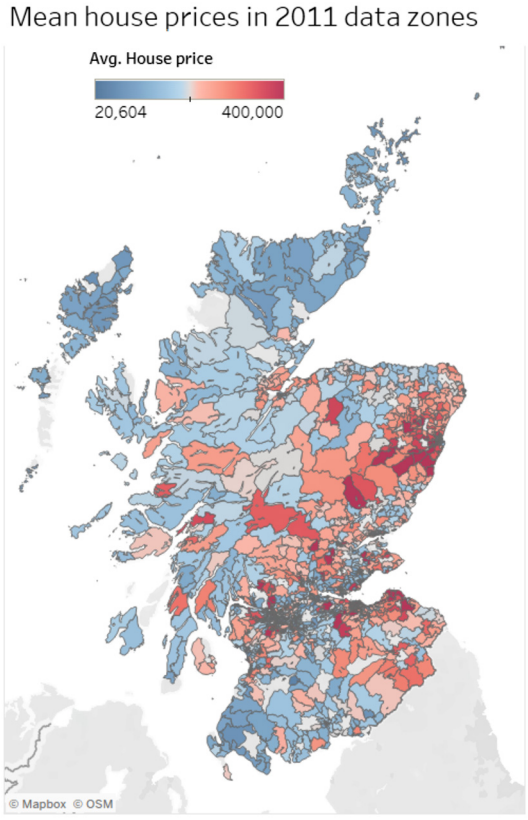


Fig. 4. Map that shows the mean house prices in 2011 data zones of Scotland in 2015. Polygons represent the data zones. The color of the polygons indicates the average house price in the data zones. Darker red highlights represent data zones with more expensive houses, while darker blue highlights data zones with less expensive houses. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

6. Create the predictive model

We use the XGBoost gradient boosting library to create the predictive model for the classification and, specifically, the scikit-learn XGBClassifier. We select "logloss" to evaluate validation data because we deal with a classification problem. Logloss indicates how close the prediction probability is to the actual value, which is 0 or 1 in our case. Higher values of log-loss mean that the predicted probability diverges from the actual value.

In order to ensure the unbiased evaluation of the model, we randomly split all data in two parts; 70% of data for training the predictive model and 30% for testing it. Towards this end, we use the train_test_split function[12] of the scikitlearn library. As a result,

[12] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html.

Table 6
Optimal hyperparameter values for the classification model. 'max_depth'; maximum depth of a tree, 'alpha'; L1 regularization term, 'subsample'; subsample ratio of the training instances, 'learning_rate'; step size shrinkage, 'n_estimators'; the number of trees, 'colsample_bytree'; subsample ratio of columns when constructing each tree, 'colsample_bylevel'; subsample ratio of columns for each level, 'scale_pos_weight'; balance of positive and negative weights.

	Parameter	Optimal value
1	n_estimators	100
2	learning_rate	0.1
3	max_depth	4
4	subsample	0.75
5	colsample_bytree	0.75
6	colsample_bylevel	0.7
7	scale_pos_weight	1.56

4510 data zones are used as training data, and 1504 as testing data. The created predictive model is then tuned using the training data in order to select optimal hyperparameters and, hence, maximize the performance of the model. Towards this end, repeated 5-fold cross-validation is used and, specifically, the sklearn Stratified K-Folds cross-validator.[13] Table 6 presents the optimal values of the hyperparameters. The scale_pos_weight is used to balance the positive and negative weights and is calculated as the number of negative instances to the number of positive instances which, in our case, is 1.56.

We evaluate the performance of the predictive model using (a) the sklearn classification report, (b) the ROC - AUC curve with confidence interval, and (c) the precision - recall AUC curve.

The sklearn classification report uses the main classification metrics including precision, recall, and f1-score. All metrics are computed based on true and false positives, and true and false negatives that are defined as:

- True Positive (TP), is the number of data zones correctly classified as having mean house prices above the average.
- True negative (TN), is the number of data zones correctly classified as not having mean house prices above the average.
- False positive (FP), is the number of data zones incorrectly classified as having mean house prices above the average.
- False negative (FN), is the number of data zones incorrectly classified as not having mean house prices above the average.

Based on the above definitions we define precision, recall (also known as "sensitivity", or "true positive rate"), and f1-score as follows:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$$

$$f1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Fig. 5 presents the classification report. There is a precision of 87% for identifying false cases (i.e., data zones that have mean house prices below the average of Scotland) and 80% for identifying true cases (i.e., data zones that have mean house prices above the average of Scotland). In addition, the recall is 87% for false cases and 81% for true cases. Finally, according to f1-score, the percentage of correct true predictions is 80% for data zones that have mean house prices above average and 87% for the rest of the data zones.

[13] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html.

	precision	recall	f1-score	support
False	0.87	0.87	0.87	1090
True	0.80	0.81	0.80	715
accuracy			0.84	1805
macro avg	0.84	0.84	0.84	1805
weighted avg	0.84	0.84	0.84	1805

Fig. 5. Results of the classification report.

Table 7
The ROC - AUC score of the predictive model along with the 95% confidence interval lower and upper limits.

ROC - AUC	Lower limit	Upper limit
0.923	0.91	0.939

We also use the ROC-AUC curve with confidence interval to evaluate the predictive model (Fig. 6a). ROC is a curve that is created after applying the predictive model to the test data and then the AUC score is computed in the holdout data. ROC curve shows the probability and AUC represents the degree of separability, i.e., how much the model is capable of distinguishing between the two classes (true and false). In our case, the higher the AUC, the better the model distinguishes between data zones with mean house prices above and below the average of all data zones. We plot the ROC curve with False Positive Rate (x-axis) against the True Positive Rate (or recall) (y-axis). The dashed line is the baseline, which represents the model when it has no predictive value. As a result, the closer the ROC curve is to the baseline, the less accurate is the prediction. The confidence interval represents the range (defined by lower and upper values) that possibly contains the mean value of the dependent variable.

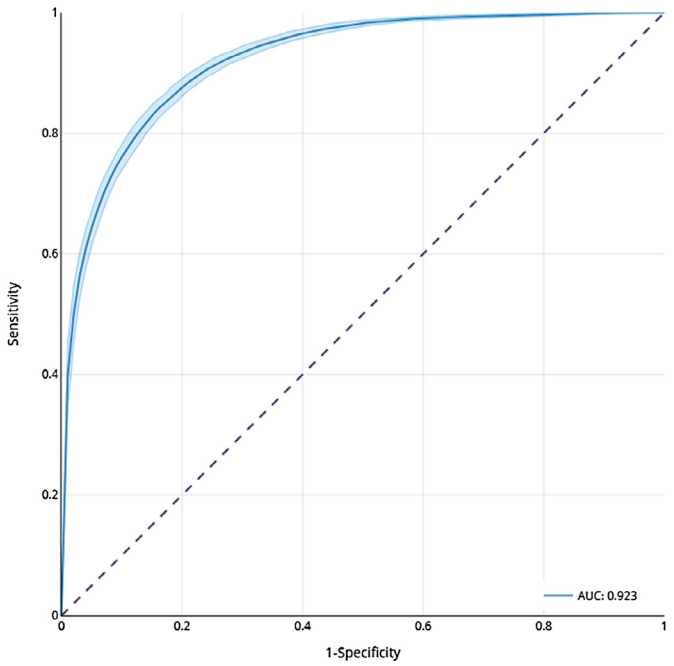
According to Table 7, the holdout AUC score is 0.923 with upper and lower bounds of confidence interval 0.91 and 0.939 respectively, which means that the predictive model is reliable.

In addition, we use the precision - recall curve (Fig. 6b) to evaluate the performance of the predictive model. The precision - recall curve is suitable for imbalanced observations between the two predictive classes since precision and recall are both based on calculations that do not make use of the true negatives but, on the contrary, are concerned only with the correct prediction of the minority class, i.e., the class with the fewer observations. In our case, the minority class regards the predictions of mean house prices of data zones that are above the average. The precision-recall curve plots the precision (y-axis) against the recall (x-axis) for various thresholds. A high area under the precision - recall curve (close to the upper right corner) means that both recall and precision are high and, hence, the predictive model returns accurate results (high precision) and the majority of the results are correctly predicted (high recall). The AUC in the precision - recall curve is 0.891 showing a reliable predictive model.

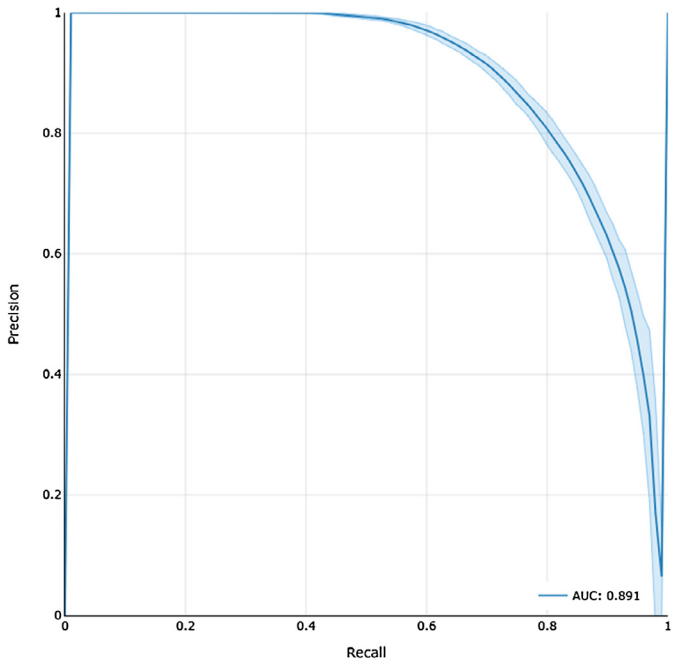
Finally, we use the XGBoost gradient boosting library to create the regression predictive model (scikit-learn XGBRegressor) that predicts the actual house prices in 2011 data zones in 2015. We use the train_test_split function to randomly split data in two parts (70% of data for training the predictive model and 30% for testing it) and tune the model using cross-validation to get the optimal hyperparameters (Table 8). The Root Mean Square Error (RMSE) for the regression model turns out to be 56692.74.

7. Explain the predictive model

This Section explains the results of the created classification predictive model. Towards this end, we employ the SHAP frame-



(a) The ROC - AUC curve. The AUC score is 0.923.



(b) The Precision - Recall curve. AUC score is 0.891.

Fig. 6. The ROC - AUC and Precision - Recall curves with 95% confidence interval.

Table 8
Optimal hyperparameter values for the regression model. 'max_depth'; maximum depth of a tree, 'alpha'; L1 regularization term, 'subsample'; subsample ratio of the training instances, 'learning_rate'; step size shrinkage, 'n_estimators'; the number of trees, 'colsample_bytree'; subsample ratio of columns when constructing each tree, 'colsample_bylevel'; subsample ratio of columns for each level, 'scale_pos_weight'; balance of positive and negative weights.

	Parameter	Optimal value
1	n_estimators	360
2	learning_rate	0.1
3	max_depth	3
4	subsample	0.75
5	colsample_bytree	0.77
6	colsample_bylevel	0.7
7	scale_pos_weight	1.56

work to globally and locally understand the decisions made by the model. In our case, global explainability is used to understand which variables mostly affect the decisions of the predictive model in overall, i.e., across all Scottish data zones, and how, whereas local explainability in specific data zones. We use the SHAP summary plot, a bar plot based on the calculated Shapley values, and dependence plots for the global explainability, and decision plots to locally explain the created predictive model.

7.1. Global explainability

In order to understand how the predictor variables affect the mean house prices in all data zones of Scotland, we first create the SHAP summary plot of Fig. 7. The SHAP summary plot is presented in a form of a beeswarm plot showing the twenty variables with the highest impact in the output of the predictive model. The plot presents (a) the predictor variables in a descending global importance, and (b) the contribution of each predictor variable in the model output for different value ranges of the variable. Each dot in the plot represents one observation, hence, in our case, one 2011 data zone. The Y-axis presents the predictor variables used to create the model in order of importance from top to bottom. The X-axis shows the Shapley values of each data zone for every variable, which indicate the probability of success; higher Shapley values mean higher probability of having expensive houses in a data zone. The color of the dots demonstrates the value of a predictor variable in a data zone. The color ranges from red to blue with red dots representing high values and blue dots low values of the variable. Large number of same color dots concentrated in close Shapley values unveil interesting patterns, which represent the contribution of a predictor variable in the model output for different value ranges of the variable. For example, in many variables red and blue dots are concentrated in opposite sites of the X-axis, meaning that high and low values of the same variable have the opposite effect in the final prediction. This is the case, for example, in the Comparative Illness Factor and Detached dwellings (ratio) variables.

According to the plot in Fig. 7, Comparative Illness Factor (CIF) and the ratio of detached dwellings in the data zones are the two features with the highest importance for the predictions of the model. Comparative Illness Factor (CIF) is an indicator of health conditions. Greater CIF values indicate poorer health conditions. The Scotland average CIF is 100, hence, data zones with values of CIF greater than 100 indicate poorer health conditions related to Scotland (and vice-versa). Data zones with poorer health conditions are represented in the summary plot with the red color and have low Shapley values, meaning that the probability for more expensive houses is low. In the same way, data zones with lower values of CIF are represented in the plot with the blue color and have greater Shapley values, hence the same probability is high. Accordingly, data zones with greater ratios of detached dwellings (red color) have probably more expensive houses related to the rest data zones of Scotland.

The SHAP summary plot also presents the distribution of effect sizes, such as the long right tails of some predictor variables. These long tails mean that specific features with a low global impor-

Fig. 7. SHAP summary plot.

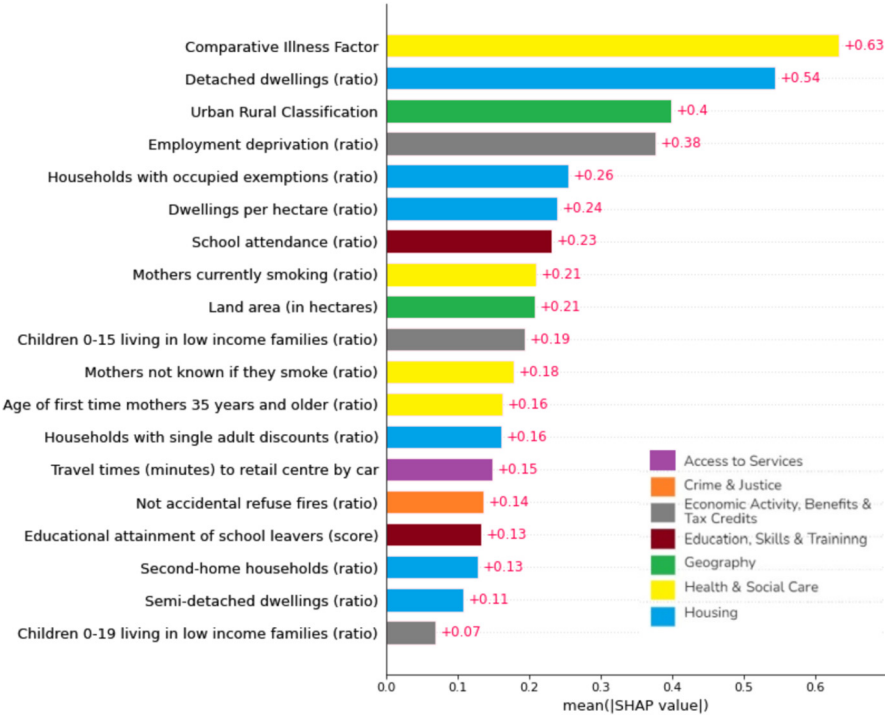
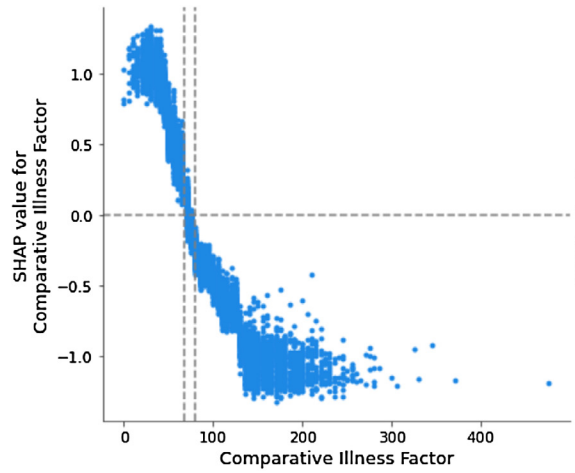


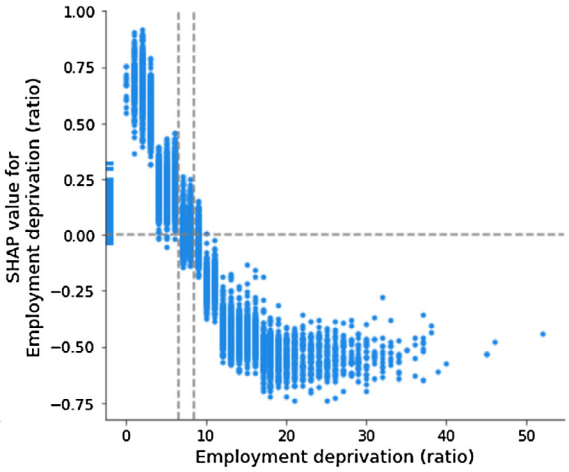
Fig. 8. Feature importance bar plot based on Shapley values.

tance can be extremely important for the house prices of specific data zones. For example, the ratio of second-home households is ranked low in the summary plot, hence it does not significantly affect house prices. However, in some irregular cases, data zones with great ratios of second-home households imply very expensive houses.

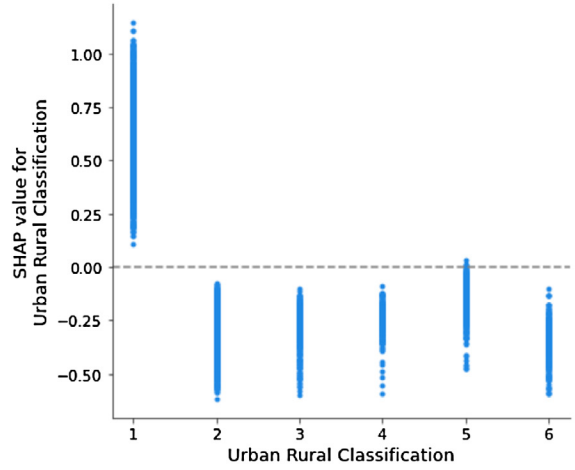
We further fine-tune the magnitude of the predictive model to determine the optimal number of variables. The predictive model with the optimal number of variables contains 19 variables. The bar plot in Fig. 8 ranks the mean Shapley values of the 19 variables. Variables are plotted on the y-axis in descending order of importance from top to bottom. The x-axis shows the absolute mean of Shapley values. The numbers on the right side indicate the respective mean Shapley values of the variable. The Figure also categorizes most important variables into themes in order to understand which themes are the most important in the prediction of the mean house prices (please refer to Table 4 for the themes and the number of variables per theme used to create the predictive model). Results show that the "Housing" theme mostly affect the predictions with six variables (31.6%), "Health and Social Care" with four important variables (21%), "Economic Activity, Benefits and Tax Credits" with three variables (15.8%), "Education, Skill and Training" and "Geography" with two variables each (10.5% each),



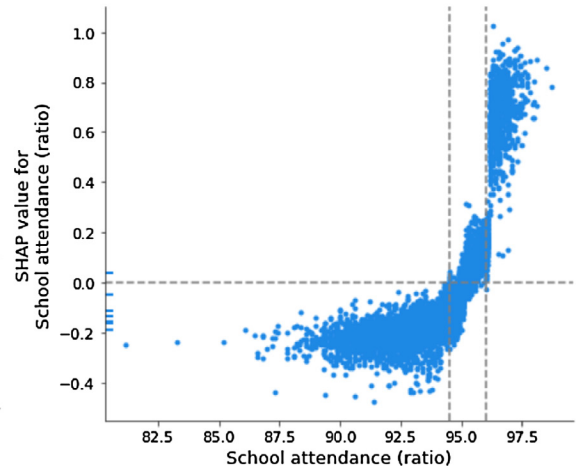
(a) CIF vs. its Shapley value



(b) Employment deprivation ratio vs. its Shapley value



(c) Urban Rural Classification vs. its Shapley value



(d) School attendance ratio vs. its Shapley value

Fig. 9. SHAP dependence plots.

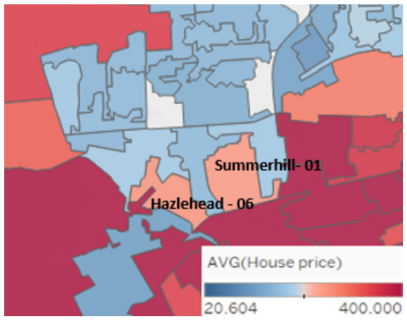


Fig. 10. The Hazlehead - 06 and Summerhill - 01 data zones.

and "Access to Services" and "Crime and Justice' with a single variable each (5.3% each).

We also use SHAP dependence plots to show the change in Shapley values while the values of a particular variable evolve. A positive Shapley value indicates that the probability for mean house price in the data zone above the average is increased as opposed to negative values which are connected with low probability for increased mean house prices. Fig. 9 presents the dependence plots of four variables vs. their Shapley values.

Specifically, Fig. 9a shows the dependence plot of CIF (x-axis) vs. its Shapley value (y-axis). The plot illustrates that there is an increased probability for mean house prices above the average of Scotland when CIF is below 70 (positive Shapley values), i.e., in richer data zones, while houses are less expensive in data zones with CIF greater than 80, which are data zones with poorer health conditions.

In addition, Fig. 9b shows the ratio of employment deprivation in data zones and the corresponding Shapley values. Data zones that have more than 9% employment deprivation have cheaper houses. In the same way, in data zones with employment deprivation less than 7% the probability for expensive houses is high.

Finally, in data zones located in large urban areas of over 125,000 people, i.e., Urban Rural Classification has value 1, (Fig. 9c) or where the attendance rate of pupils attending publicly funded schools is above 96% Fig. 9d, houses are expensive. On the contrary, data zones located elsewhere or where the attendance rate of pupils is below 94% houses are cheaper (Figs. 9c and 9d respectively).

7.2. Local explainability

This Section aims to locally explain the decisions made by the predictive model. Towards this end, we create decision plots for individual data zones.

Hazlehead - 06 and Summerhill - 01 are neighboring data zones of the council area of Aberdeen city in the north-east of Scotland

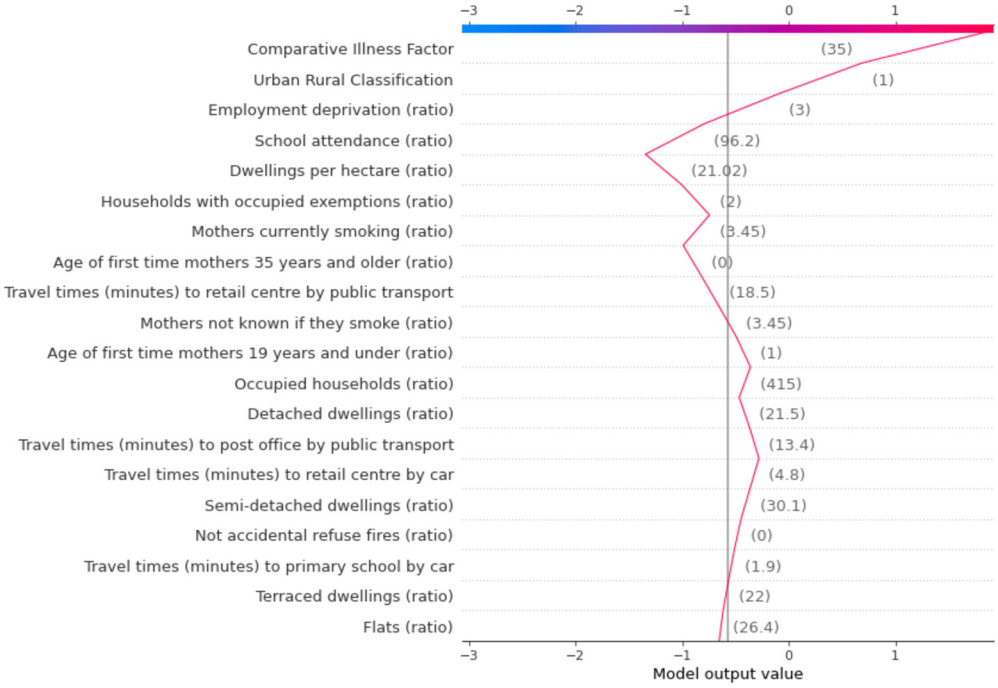


Fig. 11. Decision plot for Hazlehead - 06 data zone.

Fig. 12. Decision plot for Summerhill - 01 data zone.

(Fig. 10). Their average house prices are £257,621 and £251,658 respectively, both of them over the Scotland's mean.

Figs. 11 and 12 present the decision plots for the two data zones. The straight vertical line in the decision plots mark the models' base values, which are the average of all observations. The colored lines are the predictions and head to the left or right given the forces of the predictors. The numbers presented in the parenthesis are the variables' values. Starting at the bottom of the plot, the prediction line shows how the Shapley values accumulate from the base value to arrive at the model's final score at the top of the plot. The lines colored in blue result in final class value 0 (which indicates mean house prices below the average) as opposed to the red lines that final class value 1 (which indicates mean house prices above the average).

In both data zones the Shapley values are positive, hence, the model correctly predicted that they have average house prices above the mean of Scotland's. Specifically in Fig. 11, CIF is the most important variable for Hazlehead - 06 data zone, which agrees with the summary plot in our global explanation of the model (see Fig. 7). The value of CIF in this data zone is 35, which is lower than

70, which also indicates richer data zones according to Fig. 9a. On the contrary, although Summerhill - 01 data zone is neighboring Hazlehead - 06, CIF is not among the variables that significantly affect house prices. This is supported by the fact that the value of CIF is 75 for Summerhill - 01, which is in the range 70 to 80 that, according to 9a, indicate that they do not affect house prices.

The ratio of employment deprivation is 3% and 6% respectively in these two data zones. Both values imply more expensive houses according to Fig. 9b since they are lower than 7%. However, it is the third most important variable for the prices of houses in Hazlehead - 06 data zone, while for the house prices in Summerhill - 01 it is ranked very lower in the list of important variables. This can be explained by the fact that its value is very close to 7% which is the upper threshold for expensive houses according to 9b.

Urban Rural Classification is, as expected, important for both data zones since it indicates for both of them a large urban areas of over 125,000 people (value 1) and, also, as found in Fig. 9c, these urban areas are expected to have increased house prices.

Finally, the ratio of school attendance is, respectively, the fourth and third important variable for the house prices of the two data zones. Both of them were expected since the ratio of school attendance is 96.2% and 96.5%, and according to Fig. 9d, values above 96% indicate expensive houses.

8. Discussion

In this paper we used OGD from https://statistics.gov.scot and machine learning and created a classification model that accurately predicts whether the mean house prices in the 2011 data zones of Scotland are above the mean of Scotland. The case implemented in the paper demonstrated that linked OGD facilitate the discovery of compatible datasets, the retrieval of their data, and their integration in order to create machine learning scenarios. In addition, XAI was also used to globally and locally explain the decisions of the model providing results that can be utilized by policy makers to make better and more transparent decisions. Artificial intelligence and machine learning have nowadays become an essential issue on the agenda of governments globally. At the same time, there is a rising number of related scientific works in literature applying machine learning methods to OGD in order to predict measures of sense of community and participation (e.g., [54]), for tax recommendation (e.g., [55]), and to predict the severity of road traffic accidents [56]. However, using AI for decision support in the public sector, its consequences and potential benefits have not yet been fully explored [57]. In addition, the need for using XAI techniques for explaining individual decisions of predictive models, has only recently emerged. Although the first findings in literature suggest that XAI can help decision makers make more correct decisions, its value and feasibility need to be further investigated [58].

From the total of 59 predictor variables initially collected from https://statistics.gov.scot, we used the 19 most important ones based on their calculated Shapley values to create the predictive model. The majority of the most important variables fall into the housing theme (31.6%), followed by the health and social care variables (21%), the economic activity, benefits and tax credits variables (15.8%), education, skills and training, and geography variables (10.5% each), and access to services and geography variables (5.3% each). Although these findings are not new to literature, they are in line with it. According to previous works that explore the factors that affect house prices, apart from the physical and structural characteristics of houses, other factors that influence houses prices include the location of the houses [59,3,60], their neighborhood amenities including the accessibility to services like schools, hospitals, etc. [61,5,62], environmental characteristics such as economic status, Gross domestic product (GDP), and taxes [63,60].

9. Conclusion

This paper applies the XGBoost ensemble algorithm and XAI on linked OGD to create an accurate predictive model and explain its decisions respectively and, hence, assist house owners, buyers and investors to understand the factors that affect the house prices. Towards this end, we present a case that retrieves linked data from the Scottish official data portal to predict the probability the mean house price in 2015 in different data zones to be higher than the average price in Scotland, which is £163,478. We also predict the actual house prices in 2011 data zones.

We used SPARQL queries and retrieved 59 predictor variables coming from seven themes from the Scottish data portal. These variables were used to create the predictive model. The created classification model has Receiver Operating Characteristic (ROC) AUC score 0.923 and Precision Recall Curve (PRC) AUC score 0.891 showing a reliable model. The RMSE of the regression model is 56692.74.

In addition, we employed the Shap framework to globally and locally explained the decisions of the predictive classification model. Comparative Illness Factor (CIF), the ratio of detached dwellings, Urban Rural classification, and the ratio of employment deprivation are the variables that mostly affect the decisions of the model. Specifically, the XAI analysis showed that CIF values below 70 indicate more expensive data zones as opposed to CIF values above 80. In the same way, data zones urban areas of over 125,000 people, with ratio of pupils attendance in schools above 96%, or employment deprivation less than 7% are more likely to have expensive houses.

The local explainability of the model showed that the factors that globally determine the house prices in data zones may, in extreme cases, be less important in some data zones. An example was given were the average prices of two neighboring data zones both predicted to be over the Scotland's mean, were affected by different variables. As a result, using XAI techniques can be used to help policy makers make decisions that take into account the characteristics and factors that affect the specific geographic area (e.g., data zone).

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Law, B. Paige, C. Russell, Take a look around: using street view and satellite images to estimate house prices, ACM Trans. Intell. Syst. Technol. 10 (5) (sep 2019), https://doi.org/10.1145/3342240.

[2] C. Chwiałkowski, A. Zydroń, Socio-economic and spatial characteristics of Wielkopolski national park: application of the hedonic pricing method, Sustainability 13 (9) (2021) 1–17.

[3] K. Wongleedee, Important marketing decision to purchase condominium: a case study of Bangkok, Thailand, The Business and Management Review 9 (1) (2017) 122–125.

[4] Y. Xiao, X. Chen, Q. Li, X. Yu, J. Chen, J. Guo, Exploring determinants of housing prices in Beijing: an enhanced hedonic regression with open access POI data, ISPRS International Journal of Geo-Information 6 (11) (2017), https://doi.org/10.3390/ijgi6110358.

[5] V. Taecharungroj, Google maps amenities and condominium prices: investigating the effects and relationships using machine learning, Habitat International 118 (2021) 102463, https://doi.org/10.1016/j.habitatint.2021.102463.

[6] S. Levantesi, G. Piscopo, The importance of economic variables on London real estate market: a random forest approach, Risks 8 (4) (2020) 112.
[7] I. Gollini, B. Lu, M. Charlton, C. Brunsdon, P. Harris, GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models, Journal of Statistical Software 63 (17) (2015) 1–50, https://doi.org/10.18637/jss.v063.i17.
[8] S.C. Bourassa, E. Cantoni, M. Hoesli, Spatial dependence, housing submarkets, and house price prediction, The Journal of Real Estate Finance and Economics 35 (2) (2007) 143–160.
[9] S. Bourassa, E. Cantoni, M. Hoesli, Predicting house prices with spatial dependence: a comparison of alternative methods, Journal of Real Estate Research 32 (2) (2010) 139–160.
[10] L. Anselin, N. Lozano-Gracia, Spatial hedonic models, in: Palgrave Handbook of Econometrics, Springer, 2009, pp. 1213–1250.
[11] F. Wang, Y. Zou, H. Zhang, H. Shi, House price prediction approach based on deep learning and ARIMA model, in: 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), 2019, pp. 303–307.
[12] B. Park, J.K. Bae, Using machine learning algorithms for housing price prediction: the case of Fairfax County, Virginia housing data, Expert Systems with Applications 42 (6) (2015) 2928–2934, https://doi.org/10.1016/j.eswa.2014.11.040.
[13] A. Varma, A. Sarma, S. Doshi, R. Nair, House price prediction using machine learning and neural networks, in: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), IEEE, 2018, pp. 1936–1939.
[14] L. Hu, S. He, Z. Han, H. Xiao, S. Su, M. Weng, Z. Cai, Monitoring housing rental prices based on social media: an integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies, Land use policy 82 (2019) 657–673.
[15] Y. Kang, F. Zhang, W. Peng, S. Gao, J. Rao, F. Duarte, C. Ratti, Understanding house price appreciation using multi-source big geo-data and machine learning, Land Use Policy 111 (2021) 104919, https://doi.org/10.1016/j.landusepol.2020.104919.
[16] Y. Gao, M. Janssen, Generating value from government data using AI: an exploratory study, in: G. Viale Pereira, M. Janssen, H. Lee, I. Lindgren, M.P. Rodríguez Bolívar, H.J. Scholl, A. Zuiderwijk (Eds.), Electronic Government, Springer International Publishing, Cham, 2020, pp. 319–331.
[17] E. Kalampokis, D. Zeginis, K. Tarabanis, On modeling linked open statistical data, Journal of Web Semantics 55 (2019) 56–68, https://doi.org/10.1016/j.websem.2018.11.002.
[18] A. Deeks, The judicial demand for Explainable Artificial Intelligence, Columbia Law Review 119 (7) (2019) 1829–1850.
[19] D. Gunning, D. Aha, DARPA's explainable artificial intelligence (XAI) program, AI Magazine 40 (2) (2019) 44–58.
[20] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[21] E. Kalampokis, E. Tambouris, K. Tarabanis, Open government data: a stage model, in: M. Janssen, H.J. Scholl, M.A. Wimmer, Y.-h. Tan (Eds.), Electronic Government, Springer, Berlin, Heidelberg, 2011, pp. 235–246.
[22] B. Ansari, M. Barati, E.G. Martin, Enhancing the usability and usefulness of open government data: a comprehensive review of the state of open government data visualization research, Government Information Quarterly 39 (1) (2022) 101657, https://doi.org/10.1016/j.giq.2021.101657.
[23] E. Kalampokis, E. Tambouris, K. Tarabanis, ICT tools for creating, expanding and exploiting statistical linked Open Data, Statistical Journal of the IAOS 33 (2017) 503–514, https://doi.org/10.3233/SJI-150190, p. 2.
[24] F.S. Tseng, C.-W. Chen, Integrating heterogeneous data warehouses using XML technologies, Journal of Information Science 31 (3) (2005) 209–229, https://doi.org/10.1177/0165551505052467.
[25] S. Berger, M. Schrefl, From federated databases to a federated data warehouse system, in: Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), 2008, p. 394.
[26] L. Cabibbo, R. Torlone, A logical approach to multidimensional databases, in: International Conference on Extending Database Technology, Springer, 1998, pp. 183–197.
[27] A. Datta, H. Thomas, The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses, Decision Support Systems 27 (3) (1999) 289–301.
[28] E. Kalampokis, E. Tambouris, K. Tarabanis, Linked open cube analytics systems: potential and challenges, IEEE Intelligent Systems 31 (5) (2016) 89–92, https://doi.org/10.1109/MIS.2016.82.
[29] J.M. Perez Martinez, R. Berlanga, M.J. Aramburu, T.B. Pedersen, Integrating data warehouses with web data: a survey, IEEE Transactions on Knowledge and Data Engineering 20 (7) (2008) 940–955, https://doi.org/10.1109/TKDE.2007.190746.
[30] R. Cyganiak, D. Reynolds, The RDF data cube vocabulary: W3C recommendation, W3C Tech. Rep., 2014.
[31] A. Miles, S. Bechhofer, SKOS simple knowledge organization system reference, W3C recommendation, 2009.
[32] E. Kalampokis, A. Karamanou, K. Tarabanis, Interoperability conflicts in linked open statistical data, Information 10 (8) (2019), https://doi.org/10.3390/info10080249.
[33] Y. Freund, R.E. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, in: P. Vitányi (Ed.), Computational Learning Theory, Springer, Berlin, Heidelberg, 1995, pp. 23–37.
[34] J.H. Friedman, Greedy function approximation: a gradient boosting machine, The Annals of Statistics 29 (5) (2001) 1189–1232, https://doi.org/10.1214/aos/1013203451.
[35] Z. Peng, Q. Huang, Y. Han, Model research on forecast of second-hand house price in Chengdu based on XGBoost algorithm, in: 2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT), IEEE, 2019, pp. 168–172.
[36] Y. Zhao, G. Chetty, D. Tran, Deep learning with XGBoost for real estate appraisal, in: 2019 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2019, pp. 1396–1401.
[37] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on Explainable Artificial Intelligence (XAI), IEEE Access 6 (2018) 52138–52160, https://doi.org/10.1109/ACCESS.2018.2870052.
[38] M. van Lent, W. Fisher, M. Mancuso, An explainable artificial intelligence system for small-unit tactical behavior, in: Proceedings of the 16th Conference on Innovative Applications of Artificial Intelligence, IAAI'04, AAAI Press, 2004, pp. 900–907.
[39] J.D. Moore, Explanation in expert systems: a survey, Tech. rep., University of Southern California, Marina del Rey Information Sciences Inst, 1988.
[40] S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, T. Adams, D.E. Liston, D.K.-W. Low, S.-F. Newman, J. Kim, S.-I. Lee, Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat Biomed Eng 2 (10) (2018) 749–760.
[41] A. Laios, E. Kalampokis, R. Johnson, A. Thangavelu, C. Tarabanis, D. Nugent, D. De Jong, Explainable artificial intelligence for prediction of complete surgical cytoreduction in advanced-stage epithelial ovarian cancer, Journal of Personalized Medicine 12 (4) (2022), https://doi.org/10.3390/jpm12040607.
[42] A. Laios, E. Kalampokis, R. Johnson, S. Munot, A. Thangavelu, R. Hutson, T. Broadhead, G. Theophilou, C. Leach, D. Nugent, D. De Jong, Factors predicting surgical effort using explainable artificial intelligence in advanced stage epithelial ovarian cancer, Cancers 14 (14) (2022), https://doi.org/10.3390/cancers14143447.
[43] S. Petsis, A. Karamanou, E. Kalampokis, K. Tarabanis, Forecasting and explaining emergency department visits in a public hospital, Journal of Intelligent Information Systems 59 (2) (2022) 479–500, https://doi.org/10.1007/s10844-022-00716-6.
[44] A.B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, A.K. Mohammadian, Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis, Accident Analysis & Prevention 136 (2020) 105405, https://doi.org/10.1016/j.aap.2019.105405.
[45] M. Hind, D. Wei, M. Campbell, N.C. Codella, A. Dhurandhar, A. Mojsilović, K. Natesan Ramamurthy, K.R. Varshney, TED: teaching AI to explain its decisions, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics and Society, 2019, pp. 123–129.
[46] G. Plumb, D. Molitor, A.S. Talwalkar, Model agnostic supervised local explanations, Advances in Neural Information Processing Systems 31 (2018).
[47] J. Dieber, S. Kirrane, A novel model usability evaluation framework (MUsE) for explainable artificial intelligence, Information Fusion 81 (2022) 143–153, https://doi.org/10.1016/j.inffus.2021.11.017.
[48] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, W. Samek, Explainable AI Methods - a Brief Overview, Springer International Publishing, Cham, 2022, pp. 13–38.
[49] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red, Hook, NY, USA, 2017, pp. 4768–4777.
[50] L.S. Shapley, A value for n-person games, Contributions to the Theory of Games (AM-28) II (1953) 307–318, https://doi.org/10.1515/9781400881970-018.
[51] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence 2 (1) (2020) 56–67, https://doi.org/10.1038/s42256-019-0138-9.
[52] S.M. Lundberg, G.G. Erion, S.-I. Lee, Consistent individualized feature attribution for tree ensembles, arXiv:1802.03888 [abs], 2018.
[53] A.M. Carrington, P.W. Fieguth, H. Qazi, A. Holzinger, H.H. Chen, F. Mayr, D.G. Manuel, A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms, BMC Medical Informatics and Decision Making 20 (1) (2020) 4, https://doi.org/10.1186/s12911-019-1014-6.
[54] A. Piscopo, R. Siebes, L. Hardman, Predicting sense of community and participation by applying machine learning to open government data, Policy & Internet 9 (1) (2017) 55–75.
[55] T. Cha, Open government data for machine learning tax recommendation, in: The 21st Annual International Conference on Digital Government Research, 2020, pp. 331–333.
[56] E. Boonserm, N. Wiwatwattana, Using machine learning to predict injury severity of road traffic accidents during new year festivals from Thailand's Open Government Data, in: 2021 9th International Electrical Engineering Congress (iEECON), IEEE, 2021, pp. 464–467.

[57] D. Valle-Cruz, V. Fernandez-Cortez, J.R. Gil-Garcia, From E-budgeting to smart budgeting: exploring the potential of artificial intelligence in government decision-making for resource allocation, Government Information Quarterly 39 (2) (2022) 101644, https://doi.org/10.1016/j.giq.2021.101644.
[58] M. Janssen, M. Hartog, R. Matheus, A.Y. Ding, G. Kuk, Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government, Soc. Sci. Comp. Rev. (2020), https://doi.org/10.1177/0894439320980118.
[59] P. Bangbon, Marketing factors that affecting the purchase of condominium in Bangkok, Thailand, Psychology and Education Journal 58 (1) (2021) 4434–4438.
[60] J. Hong, H. Choi, W.-s. Kim, A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea, International Journal of Strategic Property Management 24 (3) (2020) 140–152.
[61] R.N. Belcher, E. Suen, S. Menz, T. Schroepfer, Shared landscapes increase condominium unit selling price in a high-density city, Landscape and Urban Planning 192 (2019) 103644.
[62] S. Su, J. Zhang, S. He, H. Zhang, L. Hu, M. Kang, Unraveling the impact of TOD on housing rental prices and implications on spatial planning: a comparative analysis of five Chinese megacities, Habitat International 107 (2021) 102309, https://doi.org/10.1016/j.habitatint.2020.102309.
[63] P. Boelhouwer, M. Haffner, P. Neuteboom, P. Vries, House prices and income tax in the Netherlands: an international perspective, Housing Studies 19 (3) (2004) 415–432, https://doi.org/10.1080/0267303042000204304.