Explainable Graph Neural Networks: An Application to Open Statistics Knowledge Graphs for Estimating House Prices

Areti Karamanou, Petros Brimos, Evangelos Kalampokis and Konstantinos Tarabanis

Abstract: In the rapidly evolving field of real estate economics, the prediction of house prices continues to be a complex challenge, intricately tied to a multitude of socio-economic factors. Traditional predictive models often overlook spatial interdependencies that significantly influence housing prices. The objective of this study is to leverage Graph Neural Networks (GNNs) on open statistics knowledge graphs to model these spatial dependencies and predict house prices across Scotland’s 2011 data zones. The methodology involves retrieving integrated statistical indicators from the official Scottish Open Government Data portal and applying three representative GNN algorithms: ChebNet, GCN, and GraphSAGE. These GNNs are compared against traditional models, including the tabular-based XGBoost and a simple Multi-Layer Perceptron (MLP), demonstrating superior prediction accuracy. Innovative contributions of this study include the use of GNNs to model spatial dependencies in real estate economics and the application of local and global explainability techniques to enhance transparency and trust in the predictions. The global feature importance is determined by a logistic regression surrogate model while the local, region-level understanding of the GNN predictions is achieved through the use of GNNExplainer. Explainability results are compared with those from a previous work that applied the XGBoost machine learning algorithm and the SHapley Additive exPlanations (SHAP) explainability framework on the same dataset. Interestingly, both the global surrogate model and the SHAP approach underscored the comparative illness factor, a health indicator, and the ratio of detached dwellings as the most crucial features in the global explainability. In the case of local explanations, while both methods showed similar results, the GNN approach provided a richer, more comprehensive understanding of the predictions for two specific data zones.

Keywords: linked statistical data; knowledge graphs; graph neural networks; explainable artificial intelligence; house price prediction; explainable graph neural networks

1. Introduction

Buying a house is probably one of the major decisions and financial commitments in the life of people. This decision is usually affected by prices that fluctuate due to many factors, such as changes in interest rates, economic growth, government policy, and supply and demand dynamics [1]. Changes in property prices not only reflect broader economic trends but also impact the socio-economic fabric of societies, with far-reaching consequences for home ownership and wealth distribution [2,3]. Between 2010 and 2021, the European Union experienced a surge in house prices by 37%, rents by 16%, and inflation by 17% [4]. Concurrently, the cost of construction for new residences soared by 25%, particularly since 2016. The phenomenon has been exacerbated by recent global crisis events, such as the COVID-19 pandemic and war conflict in Ukraine, resulting in unprecedented inflation levels [5] that have dramatically altered the landscape of housing prices and rent [6,7]. In such a complex environment, a realistic estimation of house prices becomes not just a theoretical exercise but an essential tool for governments, policymakers, and the private sector to form strategies and policies that reflect the ever-changing dynamics of the housing...
market. Furthermore, predicting housing prices is not only crucial for individual decision making but also for broader economic planning and policy formulation. Understanding the dynamic relationships between housing prices and socio-economic variables can inform investment decisions, fiscal policies, and urban planning initiatives [8]. This study highlights the functional relationships between residential property prices and socio-economic factors like the number of loans, unemployment levels, and market rent, showcasing the predictive power of such models over a substantial period.

In the era of accelerating digitization and advanced big data analytics, due to the growing availability of data, machine learning models have become efficient approaches for predicting house prices [9]. However, traditionally, these approaches rely heavily on socio-economic indicators to estimate house prices, which might not always yield accurate results. On the contrary, making predictions based on features with spatial context, like the presence of schools nearby, the availability of parking facilities or gas stations in the same or nearby neighborhood, and the proximity to public transport, can significantly improve the prediction accuracy [10,11] since they reflect the realistic representation of geographical regions and the connectivity between them. This results from the fact that neighboring regions may influence the house prices of each other due to their proximity and shared features. For example, the presence of schools or hospitals in a region that has high house prices might similarly affect house prices of the nearby regions. Recently, advancements in deep learning have demonstrated the superiority of approaches based on neural networks over both traditional statistical methods and conventional machine learning approaches, especially when working with geospatial data [12]. In the past few years, algorithms from graph machine learning have emerged as efficient candidates for predicting node or edge features from graph-structured data [13]. Specifically, Graph Neural Networks (GNNs) constitute a family of algorithms uniquely tailored for scenarios where the spatial representation of data must be explicitly modeled. These models have been successfully applied in various fields, such as social network analysis [14], traffic prediction [15], and recommendation systems [16], to model spatial relationships and dependencies effectively. Inspired by the widespread use and success of GNNs in these applications, we aim to apply them to house price prediction. Spatial dependencies in house prices have not been adequately modeled in traditional approaches, yet they are crucial for accurate predictions. By leveraging GNNs, we can better capture the spatial interactions and dependencies inherent in housing markets, leading to more precise and reliable predictions. Explainable artificial intelligence can be applied on top of the GNN in order to understand the decisions made by the model, improving its explainability and interpretability.

At the same time, linked data technologies have been used to facilitate the integration of open statistical data on the web. Linked statistical data are highly structured data describing statistical indicators based on a set of dimensions, including a geographical dimension and temporal dimension, and attributes [17]. As a result, linked statistical data formulate open statistics Knowledge Graphs (KGs). The structure of open statistics KGs, which already captures spatial relationships between geographical dimensions, makes them suitable for creating GNNs. Currently, open statistics KGs can be accessed as linked data by many official government data portals (e.g., the official portal for European data https://data.europa.eu/, accessed on 30 March 2024) allowing for their integration and the creation of valuable applications for governments, policy makers, enterprises, and citizens. For example, linked statistical data have been previously used to help house owners, buyers, and investors understand which factors affect and determine the prices of houses in the Scottish data zones [18].

This paper aims to create a model that accurately predicts house prices by creating a GNN using open statistics KGs. Toward this end, a case study is presented that leverages (i) three GNN variants, namely the Chebyshev Neural Network (ChebNet), Graph Convolutional Network (GCN), and GraphSAGE, and (ii) a KG modeled as linked statistical data from the Scottish data portal, to predict the probability that the average house prices across
Scotland’s “2011 data zones” are above Scotland’s total average. In order to understand the decisions of the best performed model, both global and local explainability are employed using a global surrogate model and the GNNExplainer framework, respectively.

This paper is organized as follows. Section 2 delineates the theoretical background essential to this work, specifically focusing on the theory of house price prediction, linked open government data, GNNs, and their explainability. In Section 3, the research approach utilized in the study is described, laying the foundation for the use case that follows. Section 4 is dedicated to the use case, elucidating the steps taken, including data collection, data pre-processing, and a detailed presentation of the predictive model creation and its explainability. The subsequent discussion in Section 5 compares the methods used and the results produced by this study with the methods and results of previous studies. Finally, Section 6 concludes this work.

2. Background

2.1. House Prices Prediction

The accurate estimation of the house prices in the market of real estates is of great importance for a large number of individual parties, ranging from homeowners to policy makers. Apart from the obvious supply and demand conditions, the house prices may be influenced by other factors that encompass, for example, the physical attributes of the dwellings (e.g., dwelling size, room count) [19,20], the marketing strategies employed during sale [21], environmental elements (like air quality and neighborhood safety) [22], and socio-economic indicators, including the presence of urban amenities in the area, like parks and parking facilities [19,23,24], and statistical indicators, like employment rate and population growth [25].

In the past, the task of estimating house prices was typically carried out by experienced appraisers. Nevertheless, the appraisers’ assessments could be influenced by external factors and individuals, leading to potential biases. Consequently, there has been a growing demand for using automated methods to address this issue. These methods range from conventional statistical approaches (e.g., hedonic regression, geographically weighted regression) and time series forecasting methods (e.g., autoregressive integrated moving average—ARIMA—models) [26–29] to sophisticated Artificial Intelligence (AI) solutions, including machine learning (e.g., [18,19,23,25,30–33]) and deep learning approaches (e.g., [34–37]). From the previous methods, the majority uses hedonic models and linear regression.

However, conventional statistical methods face limitations in effectively analyzing vast amounts of data, resulting in the underutilization of available information [38]. Additionally, hedonic-based regression for house price estimation has received considerable criticism for various reasons, including potential issues with fundamental assumptions and estimation (e.g., challenges in identifying supply and demand, market imbalances) [39]. Another issue with this method is that it relies on the assumption of a linear relationship between influencing factors and prices [11], as opposed to machine learning methods, which are able to explore multilevel interactions and nonlinear correlations [25].

Recently, advancements in deep learning have demonstrated the superiority of approaches based on neural networks over both traditional statistical methods and conventional machine learning approaches, especially when working with geospatial data [12]. For example, in house price prediction, neural networks can effectively capture and leverage the spatial relationships between dwellings, leading to more accurate results and promising outcomes.

Lastly, emerging big data technologies have created new types of data sources that are incorporated into the house prices prediction. These include, for example, satellite images [19,33], point of interest (POI) data, other location information from Google Maps [23], and social media data [32].
2.2. Open Government Data

Open Government Data (OGD) refers to data published by the public sector in open formats and that are freely available to be accessed and re-used by society. During the past few years, OGD have been present in the political agenda of numerous countries worldwide that aspire to improve policy-making.

The rapidly increasing velocity and diversity of public sector data have motivated the establishment of several government data portals. These portals have, as a primary objective, to provide unrestricted access to public sector data as Open Government Data (OGD), increasing their economic and social potential. Examples include the European Data Portal (https://data.europa.eu/en, accessed on 30 March 2024) and the (https://data.gov/, accessed on 30 March 2024) in the U.S., which are currently hosting more than 1,500,000 and 250,000 datasets, respectively.

Recently, the European Commission has recognized dynamic OGD, including data generated by sensors, as highly valuable data [40]. The basic feature of these data is their real-time nature, being very frequently updated. To promote the integration of such data into the development of value-added services and applications using OGD, proper access methods must be offered by OGD portals. Traditionally, OGD could be accessed from government data portals as downloadable files (e.g., CSV and JSON files). They could also be explored directly using the portals’ graphical interfaces. However, due to the growing demand for real-time data analysis, forced OGD portals are now also investing in the development of application programming interfaces (APIs). These APIs enable programmatic access to real-time data, facilitating easy and efficient analysis in real time (e.g., in [41–43]).

Nevertheless, the most prevalent types of OGD are still statistical data that are traditionally regularly but not very frequently (e.g., annually) created, published, and updated by governments and national statistical agencies. Statistical OGD are usually aggregated data monitoring various socio-economic indicators across countries, including demographic, social, and business-related metrics. Given their multidimensional nature, where measures are described based on multiple dimensions, statistical data are commonly represented using the data cube model, initially introduced for the needs of the Online Analytical Processing (OLAP) and data warehouse systems. A data cube comprises two main components [44–47]: measures that represent numerical values and dimensions that provide contextual information for the measures. For example, consider a dataset on the population at risk of poverty in European countries in 2021–2022. The data cube incorporates “risk of poverty” as a measure, described by a geospatial dimension, “European country”, and a temporal dimension, “year”. Each dimension consists of a set of distinct values; the “European country” dimension has values “GR”, “FR”, etc., while the “year” dimension has values 2022 and 2023. Optionally, there could be an additional dimension like “age group” with values “00–19”, “20–24”, “25–39”, and “50+”. The values of the dimension values can be organized in hierarchies that representing different levels of granularity. For example, the geospatial dimension might have both countries and regions as hierarchical levels while the temporal dimension might have both years and quarters. The data cube’s cells are then specified by the combination of dimension values, with each cell holding the value of the corresponding measure (e.g., the percentage of population at risk of poverty in “GR” in “2022” is “29.1%”).

2.2.1. Linked Open Government Data

Open Government Data (OGD) have a huge potential when exploited using artificial intelligence methods, including machine learning and deep learning, to bring new and fruitful insights [48]. Linked data technologies facilitate retrieving integrated OGD by defining and executing SPARQL queries. Linked data have already been adopted by numerous OGD portals (e.g., https://statistics.gov.scot/, accessed on 30 March 2024, managed by the Scottish government and the European data portal). Linked data have not only facilitated the seamless integration of data within and across different data portals.
but have also ensured the provision of high-quality data. This aspect is particularly crucial for specific types of OGD, such as statistical data, where data are described at varying levels of granularity [18]. In addition, linked data have 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 [49,50].

Connecting linked OGD from the data portals would create a knowledge graph of qualitative and fine-grained data that would facilitate data discovery and collection. In the pursuit of this objective, the existing literature has already recognized [51] and tackled [52] interoperability challenges associated with connecting data from multiple reliable and trustworthy sources. The majority of official OGD portals provide access to data using linked data through their SPARQL endpoints. In this case, OGD can be easily collected by specifying and submitting relevant SPARQL queries.

Linked data are based on the principles and philosophy of the Semantic Web. They primarily involve the publication of structured data using the Resource Description Framework (RDF), which is a W3C standard. Instead of focusing on the ontological level or inferencing, linked data use HTTP Uniform Resource Identifiers (URIs) for naming entities and concepts. This approach allows for accessing further information from looking up entities and concepts.

The naming convention in linked data follows the prefix:localname notation, where the prefix signifies a namespace URI. For instance, consider the name foaf:Person, representing the Person class of the FOAF (http://xmlns.com/foaf/spec/, accessed on 30 March 2024) vocabulary with the URI <http://xmlns.com/foaf/0.1/Person> (accessed on 30 March 2024).

The QB vocabulary [53] is a W3C standard that adopts the principles of linked data to publish statistical data on the web. By utilizing the QB vocabulary, statistical data can be effectively structured and made available as linked data, facilitating better integration and accessibility on the web. At the core of this vocabulary is the qb:DataSet. The latest serves as a representation of a data cube that comprises a collection of observations (qb:Observation), and each observation corresponds to a cell within the data cube. The data cube itself consists of three essential components:

- Dimensions (qb:DimensionProperty) that define the aspects to which the observations are applicable. Examples of dimensions include gender, reference area, time, and age.
- Measures (qb:MeasureProperty) that represent the specific phenomena or variables that are being observed and recorded within the data cube.
- Attributes (qb:AttributeProperty) that are used to convey structural metadata, such as the unit of measurement, associated with the data.

Since the dimensions with a similar context usually re-use the same values, code lists have been created and commonly used to populate the dimension values. A code list includes the URIs for all the potential values that can be used to populate the dimension. The values included in the code lists are usually specified using the QB vocabulary of the Simple Knowledge Organization System (SKOS) vocabulary [54], which is also a W3C standard.

Code lists can be hierarchical, such as the ones that can be used to populate geostatistical dimensions. Such code lists include, for example, the URIs for all geographical or administrative divisions of a country. The hierarchical relations are usually expressed using the SKOS vocabulary (e.g., using the skos:narrower property), the QB vocabulary (e.g., using the qb:parentChildProperty), or the XKOS (https://rdf-vocabulary.ddialliance.org/xkos.html, accessed on 30 March 2024) vocabulary (e.g., using the xkos:isPartOf property).

Essential concepts such as dimensions, measures, attributes, and code lists also lend themselves to reuse (i.e., employ the same URI) to enhance their discoverability. To this end, the UK Government Linked Data Working Group has delineated a set of shared concepts, drawing inspiration from the SDMX guidelines (https://github.com/UKGovLD/publishing-statistical-data, accessed on 30 March 2024). Despite not being an integral component of the QB vocabulary, these concepts enjoy widespread adoption. Noteworthy
examples of dimension concepts encompass \textit{sdmx:timePeriod}, \textit{sdmx:refArea}, and \textit{sdmx:sex}, while measure concepts include \textit{sdmx:obsValue}.

2.2.2. The Scottish Data Portal

The data used in this work are linked Open Government Data (OGD) from the Scottish OGD portal. The portal disseminates official statistics using linked data technologies. The portal currently hosts 297 linked datasets covering various societal and business aspects of Scotland classified into 18 themes.Datasets are provided at different levels of spatial granularity in Scotland starting from the level of postcodes to the level of council areas. Users can navigate through the data portal to view and retrieve data as tables, maps, and charts or download them in various formats (e.g., html, json, csv), or, alternatively, retrieve them as linked data by submitting flexible queries to the SPARQL endpoint (https://statistics.gov.scot/sparql, accessed on 30 March 2024) released by the portal.

In the Scottish data portal, datasets are made available at various spatial granularities, with a distinct hierarchy comprising sixteen (16) geographic levels. These levels span from the finest-grain postcodes to broader council areas. For instance, data zones serve as the primary geographic units for dispensing small-area statistics within Scotland. These zones were introduced subsequent to the 2001 census and underwent revisions in 2011, following the 2011 census. Comprising a total of 6976 units across the entirety of Scotland, the 2011 data zones exhibit a well-balanced distribution of resident populations, typically ranging from 500 to 1000 individuals. This population distribution ensures that each data zone contains no fewer than 375 people and no more than 1125 individuals.

The Scottish data portal offers users the ability to access and retrieve data in various formats, such as tables, maps, and charts, or to download them in formats like HTML, JSON, and CSV. Alternatively, users can use advanced queries via the SPARQL endpoint to retrieve linked data. The portal employs linked data technologies to create a unified knowledge graph, organizing datasets as data cubes and connecting them using the QB vocabulary. This allows for seamless searching across datasets and easy data combination.

Each dataset in the portal contains multiple measures, like the “Employment deprivation” dataset, which includes counts and rates of employment deprivation. Measures are identified by unique URLs that are sub-properties of the property \textit{sdmx-measure:obsValue}. The measure used in each observation is indicated through \textit{qb:MeasureType}. Dimensions such as time, geography, age, and gender are common in these data cubes. Time and geography dimensions often reuse terms from the SDMX vocabulary while age and gender dimensions have their own properties.

Dimensions are populated with values from code lists. For instance, the spatial dimension (“reference period”) might have values like calendar years, two-year intervals, or three-year intervals. These values are defined using properties like “timePeriod” for reference periods and are sometimes organized hierarchically through generalization/specialization relations in geographical hierarchies.

2.3. Graph Neural Networks

Graphs are defined as complex mathematical structures that model the relationships between various objects. They consist of vertices that represent entities and edges that represent the connections between these entities. As a result, graphs can effectively depict complex networks and systems, including social networks [14], mobility traffic networks [15,43], telecommunication networks [53], biological [56] or chemical networks [57], and so on. Deep learning has emerged as a powerful tool in artificial intelligence over the past decade, offering state-of-the-art performance in a range of fields, including speech recognition [58], computer vision [59], and natural language processing [60]. However, its application to graph data presents unique challenges [61] due to the irregular structure of graphs, their heterogeneity and diversity, and their immense size in the era of big data. In addition, the interdisciplinary nature of graphs adds another layer of complexity as it requires the integration of domain knowledge, which may not be straightforwardly in-
corporated into model designs [62]. To this end, these obstacles make it difficult to apply
traditional deep learning architectures, such as convolution and pooling, on graph data,
necessitating innovative modeling approaches. Deep learning on graphs [62] or geometric
deep learning [63] is a recently emerged field that consists of a broad range of architec-
tures, also called Graph Neural Networks (GNNs). Graph neural networks have found
widespread application across a variety of domains, from physics [64] and chemistry [65] to
computer vision [66] and natural language processing [67]. GNNs span a wide array, from
supervised to unsupervised techniques, each possessing distinct models and architectures
determined by their specific training processes and inherent convolution mechanisms,
including graph convolutional networks [68], graph recurrent neural networks [69], graph
auto-encoders [70], and graph reinforcement learning [71].

GNNs are applied on different graph structures depending on the data and use case
requirements [72]. A graph is directed (or undirected) when the direction of the connections
between entities are explicitly modeled. The types of nodes and edges determine whether
the graph is homogeneous or heterogeneous, where heterogeneous graphs consist of
different types of nodes or edges [73]. In addition, a graph is defined as dynamic when
the input features or the topology of the graph varies over time [74]. There are three
learning tasks in the context of GNNs. Node-level tasks include node classification [75] for
categorizing nodes into several classes, node regression [76] for predicting a continuous
variable for each node, and node clustering [77], aiming at defining groups of nodes
within a graph. Edge-level tasks include edge regression/classification [78] and link
prediction [79] and graph-level tasks such as graph classification [80]. Finally, from the
supervision’s method perspective, there are three categories of GNN training methods.
Fully supervised learning involves labeled data for the entire training set [81] while semi-
supervised learning involves both labeled and unlabeled nodes in the training phase using
inductive or transductive schemes for testing [82,83]. As in traditional deep learning tasks,
graph unsupervised representation learning requires only unlabeled data. For example,
graph auto-encoders apply an encoder to embed the graph into the latent representation
space, while a decoder is deployed to regenerate the graph structure [84].

In general, graph neural networks are deep learning models that use the structure
of a graph \( G = (V, E, A) \), where \( V \) is the set of nodes, \( E \) is the set of edges, and \( A \) is the
adjacency matrix, and a feature vector \( X_v \), for \( v \in V \), to learn a representation vector \( h_v \) of
a node in the latent space. To achieve this, GNNs use several neighborhood aggregation
methods to iteratively update a node’s representation by aggregating representations of its
adjacent nodes. After \( k \) layers of aggregations, a node’s representation, also named as node
embedding, contains the topological information within the node’s \( k \)-hop neighborhood.
As a high-level definition, the \( k \)-th layer of a GNN is as follows:

\[
\text{h}_v^k = \text{AGG}^k(h_u^{k-1} : u \in N(v))
\]

where \( h_v^0 \) is the vector of features of node \( v \) at the \( k \)-th layer, \( N(v) \) is the set of adjacent nodes
to \( v \), and \( \text{AGG} \) is a graph convolutional operation or a neighborhood aggregation function.

In this study, we focus on the feature representations of graph nodes for a node clas-
sification task, where the node representation \( h_v^k \) of the final layer is used for prediction.
As described in Equation (1), GNNs apply graph convolutions to generate a node’s rep-
resentation, mimicking the convolution operation applied in images [85]. There are two
categories of graph convolutions described in the literature, named spectral methods and
spatial-based methods [61,72]. Spectral methods employ graph convolutions, applying
techniques from graph signal processing [86], by transforming node representations into
the spectral domain using Fourier transformation and its variants, and spatial-based convo-
lutions apply spatial filters and neighborhood aggregation techniques to define the node’s
representation. In the following subsections, we present a brief theoretical background re-
respecting these two groups of GNN models in order to set the foundations for understanding
the results of this study.
2.3.1. Spectral Methods

Spectral methods are a class of techniques used in graph representation learning that leverage the graph signal processing theory \[86–88\]. They are built on the concept of the eigenvectors of the graph Laplacian matrix that forms an orthonormal space, facilitating the application of Fourier transformations on the feature vectors of the graph nodes, also called graph signals. The normalized Laplacian matrix of a graph and the Fourier transformation of a graph signal \(x\) are defined as follows:

\[
L = U\Lambda U^T
\]

\[
F(x) = U^T x
\]

where \(U\) is the matrix of eigenvectors and \(\Lambda\) is the diagonal matrix of eigenvalues. Graph convolutions of a graph signal \(x\) are then defined based on these transformations using a specific filter \(g_\theta\):

\[
x \ast g_\theta = U g_\theta U^T x
\]

These filters vary among different spectral convolutional graph neural networks. In the spectral convolutional neural network \[89\], these filters consist of a set of learnable parameters and are designed to deal with multi-channel graph signals. However, this architecture encounters challenges due to its dependence on the eigendecomposition of the Laplacian matrix, as well as the instability under graph perturbations and high computational complexity. On the other hand, deep convolutional graph networks \[90\] introduce smooth coefficients on spatially localized filters.

Succeeding methods such as ChebNet \[91\] and GCN \[83\] made certain approximations and simplifications to overcome these limitations. ChebNet approximates the graph filters by using Chebyshev polynomials of the diagonal matrix of eigenvalues, thus enabling the extraction of local features, also reducing computational complexity. As a result, the convolutional operation for a graph signal \(x\) can be written as follows:

\[
x \ast g_\theta = \sum_{i=0}^{K} \theta_i T_i \hat{L} x
\]

where \(\hat{L} = \frac{2}{\lambda_{\text{max}}} L - I_N\), \(\theta_i\) is a vector of Chebyshev coefficients, and the Chebyshev polynomials are defined as \(T_k = 2xT_{k-1}x - T_{k-2}x\). CayleyNet \[92\] goes further by using parametric rational complex functions to capture narrow frequency bands, with ChebNet considered as its special case.

The Graph Convolutional Network (GCN) introduces a first-order approximation to ChebNet, reducing the number of parameters to avoid overfitting, setting \(K = 1\) and \(\lambda_{\text{max}} = 2\), simplifying Equation (5) to the following form:

\[
x \ast g_\theta = \theta(I + D^{-1/2} AD^{-1/2} x)
\]

The GCN has both spectral and spatial interpretations and can be viewed as a method for aggregating feature information from a node’s neighborhood, bridging the gap between spectral-based and spatial-based approaches. There have been further improvements over the GCN, with methods like the Adaptive Graph Convolutional Network (AGCN) \[93\] and Dual Graph Convolutional Network (DGCN) \[94\] exploring alternative symmetric matrices, learning hidden structural relations, and using dual graph convolutional layers to encode both local and global structural information without the need for multiple layers.

2.3.2. Spatial Methods

Spatial-based approaches define graph convolutions based on the target node’s spatial relations similar to the convolutional operations of traditional CNNs on grids (images) \[95\]. Precisely, an image is considered as a special case of graphs with pixels represented as nodes of the grid graph. A pixel’s representation into the latent space is defined by a convolutional...
filter that computes the weighted average of pixel values of the central node as well as its adjacent nodes. Toward this direction, spatial-based approaches apply convolutional operations to define a node’s representation with several sized neighborhoods, propagating information along edges while maintaining the major local invariance of CNNs.

For example, GraphSAGE [96] is a model that generates low-dimensional node representations by sampling and aggregating features from the target node’s local neighborhood. In addition, the authors propose a framework that applies the inductive paradigm for generating node embeddings for previously unseen nodes of the graph. The graph convolution is defined as follows:

\[
h^k_v = \sigma(W^k F_k(h^{k-1}_v, h^{k-1}_u, u \in S_{N(u)}))
\]

where \(F\) is an aggregation function including the mean, LSTM, and pooling aggregator. The main novelty of GraphSAGE is a neighborhood sampling step during training, performing graph node aggregations in a batch of nodes instead of using the full neighborhood. Similarly, fast learning with GCNs [81] proposes node sampling independently for each layer, further improving the training efficiency. Other spatial-based models that employ several neighborhood sampling techniques include PinSage [97], which introduces sampling neighbors using random walks on graphs, stochastic GCNs [98], and adaptive sampling [99], which proposes mechanisms for reducing sampling variances. Apart from sampling methods, several aggregation operators are mentioned in the literature, assigning different weights to neighbors, improving the training performance and the overall accuracy of GNN models. The Graph Attention Network [100] (GAT) defines the contribution of neighboring nodes and introduces the attention mechanism by computing attention weights for determining the connection strength between two nodes. Moreover, the Gated Attention Network [101] (GAAN) performs a self-attention mechanism for each attention head. Other significant spatial-based approaches include graph isomorphism networks [102], diffusion convolutional neural networks [103], and large-scale learnable graph convolutional networks [104].

2.4. Explainability of Graph Neural Networks

Graph neural networks are adept at achieving state-of-the-art performance in representation learning for graph-structured data. However, a notable shortcoming lies in their inability to provide human-intelligible explanations for their predictions. Several methods currently exist for interpreting other types of neural networks [105], including those that generate simpler surrogate models for providing explanations [106] and those that explore interpretations of high-level features or instances impacting the model’s predictions [107,108]. However, these methods are unable to address the intricate, non-Euclidean nature of graph-structured data. In light of these limitations, much research is dedicated to devising methods for interpreting the predictions of deep graph networks. Such methods explore different aspects of the GNN models, usually with a focus on understanding the significance of input features, nodes, edges, and graph structures on the model’s predictions [109–111], providing insights for the design of new GNN-based models across various domains [112].

Perturbation-based methods employ input importance scores by analyzing the variations in predictions due to different input perturbations [113–116]. Surrogate-based methods, on the other hand, fit a simple explainable model, such as a decision tree or logistic regression, to a sampled dataset to generate explanations for the initial predictions [117–119]. Decomposition methods involve disassembling prediction scores in the last hidden layer and subsequently backpropagating these scores layer by layer to the input space. This process is conducted to compute the importance scores [120,121]. Gradient/feature-based methods utilize gradients or feature values to establish the importance of input features [111]. Lastly, model-level methods deliver high-level input-independent explanations, interpreting the general behaviors of GNNs without respect to any specific input instance [122].
The majority of these methods aim to output local explanations for a single instance identifying important parts of data that contribute to the model’s prediction. Local explainability is a crucial aspect of understanding complex machine learning models. It refers to the ability to explain the prediction for a specific instance, rather than explaining the model’s global behavior. Currently, most research efforts in GNN interpretability have primarily focused on instance-level explanation techniques, referred to as local explainability [123]. For example, a significant work in this area, GNNExplainer [113], defines an explanation in a form of a sub-graph that contains a subset of nodes and node features that contribute to the prediction. GNNExplainer is a widely used explanation method for local explainability in GNNs that operates on a perturbation-based principle. It works by identifying the most critical edges or nodes contributing to a particular prediction by creating a binary mask over the graph’s adjacency matrix. The mask is learned through gradient descent by maximizing the mutual information between the model’s predictions on the masked graph and the original prediction. This process ultimately reveals the minimal sub-graph that most significantly influences the prediction of interest. The explainer performs this process by gradually learning which edges and nodes are most important for each prediction, allowing it to mask out the parts of the graph that are not important. The final output is a sub-graph that maintains the most relevant nodes and edges for the prediction, providing an interpretable local explanation for the prediction made by the GNN. On the contrary, global explanation methods provide instance-independent, model-agnostic, and high-level interpretations of the overall behavior of the black box models [124]. However, the global understanding of GNNs is yet an unexplored area while studies on benchmarking graph explainers have focused on instance-based methods [125]. A few recent efforts toward global explainability in GNNs include generative-based techniques such as XGNN [122] and GNNInterpreter [126], GLGExplainer [127], which extends local explanations of PGE explainer, and global counterfactual explainer [128], which employs a selection of the top k counterfactuals using random walks. Furthermore, a common method that provides global explainability for complex machine learning and deep learning models is the creation of a simpler, surrogate model. A global surrogate model is an interpretable model constructed to explain the predictions of a more complex ‘black box’ model. This approach allows us to interpret the surrogate model, thereby making conclusions about any intricate model, including GNNs, without delving into its complex internal mechanisms. The surrogate model should be both interpretable and capable of mimicking the predictions of the GNN model. The primary goal is to closely approximate the GNN’s prediction function \( f \) with an interpretable surrogate model prediction function \( g \). Any interpretable model can be employed for the function \( g \).

The most prominent examples of interpretable models include decision trees [129], which can provide global interpretations displaying the entire decision tree, and linear models such as logistic regression and ridge, which can incorporate a set of weights per feature, providing global interpretations in the form of feature importance [130]. Other simple interpretable models include naïve Bayes models, k-Nearest Neighbors (k-NN) models, NNs with a single layer, and Generalized Linear Models (GLMs). The selection of the appropriate surrogate model depends on various factors, such as the need for interpretability, model complexity, and the nature of the underlying data. In the context of explaining a complex model like a graph neural network for node classification, the choice of logistic regression as a surrogate model can be justified by its emphasis on providing more interpretable results compared to other methods. Logistic regression’s linear nature allows for a clear interpretation of the relationship between input features and the output predictions, making it suitable for global explanations. On the other hand, decision trees, while capable of explaining predictions through a hierarchy of decisions, may not offer the same level of interpretability as logistic regression when it comes to understanding the impact of individual features on the model’s outcomes. Therefore, for scenarios where transparency and straightforward interpretations are crucial, logistic regression can be a more appropriate choice as a surrogate model. However, a notable limitation of linear
surrogate models is the linear assumptions made by these models. Logistic regression assumes a linear relationship between the independent variables and the log-odds of the dependent variable. This can limit its flexibility to capture complex, non-linear relationships in the data compared to other more complex surrogate models, such as decision trees and k-NNs.

The surrogate model prediction function $g$ of logistic regression is defined as follows:

$$
g = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}
$$

where $x_i$ represents features and $\beta_i$ represents the coefficients corresponding to these features. The higher the absolute value of the coefficient, the more significant the feature is in predicting a specific class. As a result, the coefficients of the logistic regression model serve as a measure of feature importance, providing insight into the original GNN model predictions.

3. Research Approach

In order to fulfill the objectives of this work, we use four steps (Figure 1).

Figure 1. Research approach of this work.

(1) Collect data (Figure 2). This work utilizes linked statistical data from the Scottish Open Government Data portal. Toward this end, multiple SPARQL queries were submitted to the SPARQL endpoint provided by the data portal. Specifically, the first query was applied to find all compatible datasets in the Scottish data portal (i.e., with the same year of reference and granularity level of the geography dimension) that measure ratio, percent, or score, resulting in 30 datasets. Then, using various years, we repeatedly submitted a second SPARQL query to determine which year had the greatest amount of compatible variables. Year 2015 was the outcome. The third SPARQL query was then submitted to retrieve the final list of datasets that will be used to obtain the statistical indicators. This query searches for datasets that measure ratio, percent, mean, or score (rank) values regarding 2011 data zones and for years 2015, 2014–2015, or 2014–2016. It resulted in 16 datasets. The statistical indicators were then retrieved by manually locking the values of the dimensions of the datasets, resulting in 60 indicators. Finally, an SPARQL query was submitted to retrieve the final values of the 60 indicators for each data zone and for the selected year. A detailed presentation of the method used to retrieve data and the SPARQL queries, as well as descriptive statistics of the dataset, can be found in [131].

Figure 2. The flowchart for the data collection step.

(2) Pre-process data (Figure 3). This step transforms the integrated statistical indicators into a geo-centric knowledge graph that is suitable for being used by Graph Neural Network (GNN) algorithms to predict the house prices in Scottish “2011 data zones”. Toward this end, data zone records with null values in every feature are initially removed. The remaining data are then formatted in a way that is centered on the “2011 data zones”. To achieve this, a geographically focused sub-graph of the original linked dataset was created. The Scottish “2011 data zone” are the central nodes of the sub-graph, each of them connected to the
values of the associated features. An illustration of this transformation is shown in Figure 4. 
More details are presented in Section 4.2.

(3) Predict house prices. This study leverages graph modeling for the prediction of house prices. Three distinct variants of GNNs are employed, each representative of the two key methodologies concerning graph convolutions, as detailed in Sections 2.3.1 and 2.3.2 of this paper: the Chebyshev Neural Network (ChebNet), Graph Convolutional Network (GCN), and GraphSAGE. The problem is formulated as a node classification task for classifying Scottish “2011 data zones” into two categories: above the mean house price of all data zones or below the mean house price of all data zones. The training/validation/test split of this dataset is aligned with the fully supervised scenario. To this end, all labels of the training examples are used for training following the implementation of [81], with a split ratio of 0.6/0.2/0.2, respectively. The created models are evaluated and then compared with each other. They are also juxtaposed with the model created in a previous work [18], which utilized the XGBoost machine learning algorithm and a straightforward Multilayer Perceptron (MLP), based on the same dataset.

(4) Explain the model. To facilitate an in-depth understanding of the mechanisms behind the model’s predictions, both local and global explainability are used to provide comprehensive insights into the model’s behavior.

Global explainability is tackled through the implementation of a simple surrogate model using logistic regression, chosen for its interpretability and simplicity. This model is trained on the predictions of the superior performing GNN based on the results of the previous step, which is GraphSAGE. Specifically, the model is trained on the original features in combination with the predicted probabilities of the binary classification outcomes from the GraphSAGE model. The subsequent calculation of feature importance is based on the coefficients of the logistic regression. It is assumed that the magnitude of the coefficients inherently determines the relative importance of each feature for the GNN’s predictions. Therefore, these coefficients form the basis for interpreting the GraphSAGE’s prediction behavior on a global level.

Local explainability, on the other hand, is addressed through the application of a well-established method, GNNExplainer [113]. Toward this direction, two neighboring data zones are selected, both with house prices above the mean. For each data zone, local feature importance is computed, determining the most critical features that influence
the model’s prediction for the specific node. Additionally, an explanation sub-graph is visualized, highlighting the most significant nodes, edges, and features that strongly impact the prediction. Interestingly, these two data zones have previously been examined in another study [18] using the XGBoost model and the SHAP explainability framework. This previous research offers a valuable benchmark, allowing for the comparison of results, not merely on the level of predictive performance but also in terms of explainability and feature importance.

4. Using Explainable Graph Neural Networks to Predict the House Prices in Scottish Data Zones

This section presents the implementation of the case study based on the steps described in the methodology of Section 3, namely, (i) Collect data, (ii) Pre-process data, (iii) Predict house prices, and (iv) Explain the model.

4.1. Collect Data

The statistical data used in this work were retrieved as linked data from the official OGD portal of Scotland. The data were retrieved from sixteen (16) datasets of the OGD portal classified into seven categories, including health and social care, housing, and crime and justice, resulting in sixty (60) statistical indicators (see Table A1 in the Appendix A). The majority of the indicators are “Crime and Justice” data (22 indicators or 37.2%), followed by “Housing” indicators (13 indicators or 22%). Excluding the Comparative Illness Factor (CIF) and urban rural classification, which are integer and categorical variables, respectively, the rest of the indicators are numeric.

A total of 6976 observations were extracted using the SPARQL queries. Each observation refers to a Scottish “2011 data zone” accompanied by its associated statistical indicators. This quantity of data zones represents an 86.2% coverage of the entire collection of “2011 data zones” within Scotland. The reason for this disparity is that certain data zones that lack values for one or more statistical indicators are not included. The main year of reference is 2015, while, for indicators pertaining to two- or three-year spans, we have chosen 2014–2015 and 2014–2016 as the designated reference periods. A small part of the data (1.4%) are null values. Comprehensive descriptive analysis of these indicators can be found in our previous work in [131].

The problem’s dependent variable is the mean house price, which ranges from GBP 20,604 (Cumbernauld Central, Glasgow) to GBP 1,244,910 (Leith (Albert Street)–03 in the city of Edinburgh) across all data zones. The average cost of a house across all data zones is GBP 163,478; the mean price of houses is higher in 39% of the “2011 data zones”. Therefore, determining whether a Scottish “2011 Data Zone”’s average house price is (a) over or (b) under GBP 163,478 is the classification problem that this study addresses. Finally, there is a small imbalance in the data.

4.2. Pre-Process Data

The Scottish statistical indicators were retrieved as linked data. The original dataset included 782 data zone records with null values in every feature. We removed these records, resulting in 6014 data zone records.

To facilitate the development of graph neural networks for house price prediction, a sub-graph was extracted from the retrieved dataset that centers on geographical attributes, particularly emphasizing the Scottish “2011 data zones”.

In Figure 4, an example of the initially retrieved graph with the Scottish statistical data is presented. The graph comprises four observations derived from two datasets (two observations per dataset). All observations pertain to 2015. The observations of dataset 1 describe “Comparative Illness Factor” (CIF), which is an indicator of health conditions, for two data zones, namely “City Centre West 01” and “City Centre East 04”. The two data zones are neighboring areas within Aberdeen, a city in North East Scotland and the third most populous city in the country. The value of CIF is 55 and 95 for the “City Centre West
01” and the “City Centre East 04” of Aberdeen, respectively. Similarly, the observations of dataset 2 describe the percentage of employment deprivation in the same data zones, which is 6% in the “City Centre West 01” and 9% in the “City Centre East 04”.

In Figure 5a, the part of the geo-centric graph that will be used for constructing the GNNs is selected. Only nodes with information regarding the “2011 data zones”, the value of the measure of the statistical indicator, and the values of additional properties (e.g., gender, age, etc.) are required. All other information, such as the dataset or the observations, is excluded. The final sub-graph is presented in Figure 5b. The edge connecting the data zone nodes represents the neighboring relationship between the data zones.

![Figure 5. (a) The nodes of the Scottish linked statistical indicators graph that will be used for constructing the GNNs. (b) Part of the final graph after the transformation of the linked data graph.](image)

The final graph comprises a total of 6014 interconnected nodes (or data zones), with 20,521 edges representing the adjacency of the respective data zones (Figure 6). Consequently, there is a connection between each data zone and its adjacent data zones. The hue of the data zones signifies the mean house price within its locale, ranging from deep blue to intense red as the mean house price ascends, thereby offering a visual gradient of housing costs.

![Figure 6. Visualization of a subset of the final graph that will be utilized for node classification using graph representation learning. Each node corresponds to a data zone and edges connect the centroids of adjacent regions. Data zones colored in shades of blue indicate regions with lower mean house prices, while red indicates higher mean house prices.](image)
4.3. House Price Prediction with Graph Neural Networks

Three GNN variants are implemented, as explained in Section 3, namely, Graph Convolutional Network (GCN), Chebyshev Neural Network (ChebNet), and GraphSAGE. The three GNN variants are compared against an XGBoost model that has been previously tested on the same dataset and against a Multilayer Perceptron (MLP) that acts as a per-node baseline classifier that does not incorporate the underlined graph structure.

The study utilizes an undirected, attributed graph, wherein each node incorporates a 59-feature vector. This graph comprises 6014 nodes and 20,521 unweighted edges. Prior to the implementation of the experiments, all features are normalized to a 0–1 scale to ensure consistent comparability across various feature ranges. The implementation details of the experiments are the following. All networks consist of two layers followed by a Relu non-linear activation function, and the final layer is used for binary classification followed by a logistic sigmoid activation. The learning rate of all networks is 0.01 and the dimension of hidden units is set to 32, while the dropout rate is 0. Furthermore, the Adam optimization method [132] is selected for all GNN models during training. Early stopping was implemented to mitigate the potential of model overfitting. If no improvement in validation accuracy was observed over 20 consecutive epochs, the training process was terminated. The cross-entropy loss function was adopted to assess the performance of all neural networks.

The original GCN is modified to the inductive setting following the work of [81]. On the contrary, GraphSAGE applies neighbor sampling and aggregation compared to the GCN that aggregates features from all neighbors. To this end, the mean aggregator is selected for GraphSAGE, with a neighbor sampling size set to 20 and 10 for each layer. In addition, for the Chebyshev spectral graph convolutional operator, the size of the Chebyshev filter K is set to 2.

Figure 7 shows the test and validation learning curves of the three GNN variants and MLP. Notably, GraphSAGE exhibited the highest accuracy scores, both in test and validation scenarios, from the very beginning of the training process. This suggests a promising level of learning efficiency and model stability in GraphSAGE. GraphSAGE’s early high accuracy scores can be attributed to its robust learning mechanism, which effectively harnesses the features of the local neighborhood of each node. The capacity to employ sampling to aggregate information from the node’s neighborhood appears to confer an initial advantage to GraphSAGE over the other architectures in the test and validation plots. Moreover, a high initial accuracy suggests that GraphSAGE requires fewer epochs to reach an optimal or near-optimal performance, leading to reduced training times and resource expenditure, while all other models are trained for longer epochs.

![Learning curves](image1)

**Figure 7.** Learning curves of the three different GNN variants and MLP. (a) Test learning curves. (b) Validation learning curves.

In addition, Figure 8 shows the training runtimes for the different approaches. All training times are comparable, with the GCN being the slowest and GraphSAGE the fastest due to its ability to sample and aggregate node information instead of aggregating from all nodes (as all other GNN variants do).
To validate the representational power of GNNs to incorporate the spatial dependencies among data zones, comparative evaluation results are summarized in Table 1. Moreover, to ensure the reliability and robustness of the performance comparison, each model was executed 30 times. The mean values of the performance metrics across these runs are displayed in Table 1. The evaluation involves common classification metrics, including accuracy, precision, recall, F1 score, and Area Under the ROC Curve (AUC-ROC) score. The evaluation metrics of the XGBoost model have been drawn from the prior work in [18].

Table 1. The averaged test prediction results for the supervised node classification task for the GNN variants, MLP, and XGBoost ($p < 0.05$).

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>ROC-AUC</th>
<th>Epochs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GraphSAGE</td>
<td>0.876</td>
<td>0.876</td>
<td>0.876</td>
<td>0.876</td>
<td>0.93</td>
<td>68</td>
</tr>
<tr>
<td>GCN</td>
<td>0.852</td>
<td>0.852</td>
<td>0.852</td>
<td>0.852</td>
<td>0.91</td>
<td>112</td>
</tr>
<tr>
<td>ChebNET</td>
<td>0.847</td>
<td>0.847</td>
<td>0.847</td>
<td>0.847</td>
<td>0.91</td>
<td>103</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.840</td>
<td>0.850</td>
<td>0.840</td>
<td>0.840</td>
<td>0.92</td>
<td>-</td>
</tr>
<tr>
<td>MLP</td>
<td>0.827</td>
<td>0.832</td>
<td>0.827</td>
<td>0.827</td>
<td>0.90</td>
<td>72</td>
</tr>
</tbody>
</table>

To validate the differences in performance between the models, paired $t$-tests were conducted between all pairs of models. This statistical significance testing was essential to confirm that the observed differences were not due to random chance but were statistically significant. The results of the paired $t$-tests, detailed in Appendix A Table A2, show that the differences in performance metrics between all models are statistically significant ($p < 0.05$). The GraphSAGE model achieves the highest accuracy, precision, recall, and F1 score among all models, coming in at 0.876 for each of these metrics, while also having an AUC-ROC value of 0.93. Additionally, GraphSAGE accomplishes these results in 68 epochs, indicating a relatively efficient learning process compared to the other graph-based models like GCN and ChebNET, which required 112 and 103 epochs, respectively. The GCN
and ChebNET models perform similarly on accuracy, precision, recall, and F1 score, each achieving around 0.85, but their AUC-ROC values indicate some difference in model performance. The GCN’s and ChebNet’s AUC-ROC values are 0.91, indicating a slightly better performance at distinguishing between the two classes. Considering XGBoost and the MLP, both methods fall short in comparison to the graph neural network models. Although XGBoost has an AUC-ROC score of 0.92, which is higher than the GCN, its overall accuracy, precision, recall, and F1 score are lower than all three GNN models. The MLP has the lowest performance across all metrics, despite having a competitive AUC-ROC score of 0.90. Figure 9 depicts all the precision-recall and receiver operating characteristic curves for all GNN variants and the MLP.

Figure 9. Cont.
In addition, Figure 10 depicts the classification metrics further by displaying the results for each class (0 and 1) separately. In general, the GraphSAGE model outperforms other models across most of the metrics and for both classes, demonstrating its strength in the binary classification task. This is particularly noticeable in the case of accuracy, where it consistently achieves top results. Its superiority extends to the recall for class 1, F1 score for class 0, and precision for class 1. However, there are some metrics where other models exhibit stronger performance. Interestingly, the MLP model surpasses other models in terms of recall for class 0. This suggests that the MLP, while not the best overall performer, is particularly adept at identifying true positives in the instances that are actually of class 0. Similarly, XGBoost shows effectiveness in the precision metric for class 0, indicating that when it predicts an instance to be class 0, it is likely to be correct. It also achieves higher scores for F1 class 1, demonstrating its balanced performance between precision and recall for instances of class 1. The figure also reveals some of the models’ weak points. For instance, the GCN appears to struggle with precision for class 0, where it ranks last among the models. This suggests that it has a higher rate of false positives when predicting class 0, which could be a significant drawback if precision in identifying this class is a priority.

Figure 10. Visual comparison of the performance metrics precision, recall, F1 score, and accuracy of the three GNN variants, XGBoost, and MLP. The metrics are evaluated separately for the entire dataset (‘All’), instances above the mean house prices threshold (‘Class 1’), and instances below that threshold (‘Class 0’).
Figure 11 shows the UMAP (Uniform Manifold Approximation and Projection) projection of the node embeddings learned by the top performing model GraphSAGE into a 2D space. The UMAP visualization shows a distinct separation between the two classes, which indicates that the model has learned embeddings that can effectively distinguish between the classes.

4.4. Explainability

In order to better understand the decisions of the prediction models, we use both global and local explainability. Both methods are applied to the most accurate model, i.e., GraphSAGE, based on the results of the previous sections.

4.4.1. Global Explainability

In this study, a logistic regression model is chosen as surrogate to provide global explainability. The predicted probabilities produced by the GraphSAGE model along with the initial features are used to train the logistic regression. The higher the absolute value of the coefficient, the more significant the feature is in predicting whether a house price is above or below the mean price. Each provided coefficient represents the change in the log-odds of the target variable given a one-unit change in the predictor, all other variables being held constant. Figure 12 depicts feature importance as determined by the coefficients of the surrogate model. Each bar corresponds to a feature used in the GNN model, and the length of the bar signifies the magnitude of the feature’s impact on the model’s predictions. The direction of the bar indicates the polarity of the coefficient, i.e., whether the feature contributes positively or negatively to a house price being above the mean price.

Notably, several indicators are negatively associated with a data zone having house prices above the average. For example, “Comparative Illness Factor” (CIF) shows the largest negative coefficient, suggesting that data zones with higher values for the CIF tend to have lower mean house prices. Similar trends can be observed for features such as “Households with single adult discounts (ratio)”, “Terraced dwellings”, and “Mothers currently smoking (ratio)”. On the other hand, factors that increase the likelihood of a data zone having house prices above the mean are also identified. The most substantial of these is the ratio of “Detached dwellings”, suggesting that regions with a higher proportion of detached dwellings are more likely to have higher house prices. Other significant positive factors include the proportion of “Mothers never smoked”, “Occupied households ratio”, “School
attendance ratio”, and “Flats ratio”. This could reflect the impact of education and healthier living environments on house prices. Coefficients for travel times to different services (like retail centres, schools, and GP surgeries) by car and public transport are mostly negative, suggesting that longer travel times might be associated with lower house prices. However, these coefficients are small, implying that these factors might not be influential in determining whether house prices are above or below the mean. Features related to fires show both positive and negative associations with house prices. For instance, “Accidental outdoor fires” and “Vehicle fires” have positive coefficients, indicating a slight increase in the likelihood of house prices being above the mean, while “Other primary fires” and “Accidental refuse fires” are negatively associated.

Figure 12. Global feature importance based on the logistic regression surrogate model. Each bar signifies the magnitude of the feature’s coefficient in the logistic regression model, which is interpreted as its importance in the GNN’s predictions.

4.4.2. Local Explainability

The GNNEXplainer is applied to two specific nodes corresponding to data zones S01006552 or Hazlehead-06 and S01006553 or Summerhill-01. Hazlehead-06 and Summerhill-01 are adjacent data zones located within the council area of Aberdeen city, in the north-eastern region of Scotland. The average house price in these zones is GBP 257,621 and GBP 251,658, respectively, both of which are higher than the average house prices across Scotland. The GraphSAGE model correctly predicted them as having house prices above the mean.

To understand and interpret the model’s predictions for these two data zones, the explanation sub-graph is visualized. Figures 13 and 14 depict the explanation graphs of Summerhill-01 and Hazlehead-06 data zones, respectively, illustrating the most influential nodes and edges contributing to this node’s prediction. Each node represented in this sub-graph is characterized by its most important feature. The sub-graph also depicts the structural interdependencies between the target node and its neighboring nodes. The edge opacity in the explanation sub-graph signifies the importance of each connection, with more opaque edges corresponding to stronger influential relationships.
Figure 13. Visual representation of the explanation sub-graph for data zone Summerhill-01 (S01006553), highlighted in red. Each node in the sub-graph is accompanied with a feature and its importance score, determined as the most influential by GNNExplainer for predicting the class of the target node. The opacity of the edges reflects their respective importance in this prediction process.

It is noted that all nodes within a one-hop distance from the target node have highly influential connections for both data zones, suggesting that the immediate neighborhood of a node plays a crucial role in the model’s predictions. It is observed that not all regions contribute equally to the prediction outcome. In the case of Summerhill-01 (Figure 13), data zones like S01006704, S01006705, S01006702, S01006717, S01006552, and S01006722 appear to have less influence on the prediction of the target node, all of which display house prices that fall below the mean. Furthermore, the “Employment deprivation” (ED) feature appears several times across the nodes, suggesting that it is a significant predictor for the target node. Similarly, “Comparative Illness Factor” (CIF), “Other building fires” (OBF), and “Urban Rural Classification” (URC) also come up multiple times across nodes, suggesting their significant role in predicting the target variable.

Figure 15 depicts the total feature importance for data zone Summerhill-01 as the sum of the node mask values across all nodes for each feature. This plot aggregates the importance of each feature across all nodes of the explanation sub-graph, offering a comprehensive understanding of the relative importance of features. As a result, “Urban Rural Classification” is ranked first in terms of importance, followed by “Comparative Illness Factor”, “Employment Deprivation”, “Educational Attainment of School Leavers”, and “Detached Dwellings”. In this case, the local interpretation for classifying this specific region as an area likely to have house prices above the mean aligns well with the global interpretation as presented in Section 4.4.1. It is observed that features like “Urban Rural Classification” and “Comparative Illness Factor”, which hold high importance on a global scale, also play a substantial role in driving the local prediction for this particular region.
Figure 14. Visual representation of the explanation sub-graph for data zone Hazlehead-06 (S01006552), highlighted in red. Each node in the sub-graph is accompanied with a feature and its importance score, determined as the most influential by GNNExplainer for predicting the class of the target node. The opacity of the edges reflects their respective importance in this prediction process.

Figure 15. Feature importance of the top 20 features that play a crucial role in explaining the prediction made by GraphSAGE for the data zone Summerhill-01. The total importance score is based on the sum of the node masks (obtained during model explanation) across all nodes for each feature.

Finally, in the case of Hazlehead-06, Figure 14 depicts stronger relations of the target node with the one-hop neighbors, as well as data zones S01006703 and S01006557. “Flats” (F), “Accidental Refuse Fires” (ARF), “Occupied Households”, and “Households with single adult discounts” (HWSAD) are features that prominently appear across multiple nodes, suggesting their pivotal role in forecasting whether house prices in Hazlehead-06 will surpass the mean value.

However, as seen from the feature importance plot in Figure 16, other features like “Comparative Illness Factor” and “Dwelling fires” have high aggregate importance scores. While the explanation sub-graph provides valuable insights into the structural dependencies and reveals localized feature influences, the feature importance plot captures the collective importance of each feature in the model’s predictive behavior by aggregating the importance scores of all features across all nodes of the explanation graph. Furthermore, it is noteworthy that key features such as “Comparative Illness Factor”, “Flats”, “Mothers who are Former Smokers”, and “Occupied Households” appear in both global and local importance plots. The consistency indicates that despite their different methodologies...
Technologies 2024, 12, 128

and assumptions, both models perceive these particular features as having a significant influence on the prediction of house prices in Hazlehead-06.

![Feature importance of the top 20 features](image)

**Figure 16.** Feature importance of the top 20 features that play a crucial role in explaining the prediction made by GraphSAGE for the data zone Hazlehead-06.

5. Discussion

This study utilizes graph neural networks to forecast housing prices across Scotland’s “2011 data zones”. House price prediction, a research area that is rapidly gaining traction in the current academic literature [18, 23, 25], is a problem intricately intertwined with spatial dependencies [27, 133]. Recognizing that house prices of a region are not solely determined by its inherent characteristics, such as health and social care indicators, economic statistics, or housing details, but also by the socio-economic context of its surrounding regions, it becomes evident that models must incorporate these spatial interdependencies. Spatial dependencies play a critical role in house price prediction as they account for the influence of neighboring areas on the target region. For example, a prosperous area can have a positive impact on the house prices of adjacent zones due to spillover effects. Conversely, regions with high crime rates or poor infrastructure can adversely affect the house prices of nearby areas. This interconnected nature of regions highlights the necessity for models like GNNs that can effectively capture and utilize these spatial dependencies to improve prediction accuracy. GNNs, by their nature, are uniquely suited to handle this spatial context due to their inherent ability to model relationships between entities, hence providing an innovative approach to tackle house price prediction.

However, GNNs, like many other sophisticated machine learning models, are often referred to as “black box” models due to their opaque decision-making processes. The explainability of such models remains a largely unexplored area. While numerous explainability techniques have been proposed, many are unsuitable for GNNs because they fail to consider the nature of graph-structured data, while most applications in GNN explainability focus on areas like biology and chemistry [109, 134] rather than house price prediction. Moreover, another shortcoming in the existing body of GNN research is the absence of global explanations, with most interpretability efforts focusing on instance-level explanations [113, 114, 117]. This narrows the scope of the interpretability, preventing comprehensive insights into the overall decision-making process of the GNN. In this study, we tackle this limitation head-on by deploying an interpretable surrogate model for the “black box model” to provide a global explanation for its predictions.

Finally, this work compares the outcomes of the graph-based approach with a previously conducted study using the XGBoost machine learning algorithm and the SHAP explainability framework [18].

From the results, it can be observed that the GNNs produced accurate predictions for classifying data zones above or below the mean house prices. Comparatively, the GNN
variants (ChebNet, GCN, GraphSAGE) outperformed the XGBoost algorithm utilized in the prior research. Notably, GraphSAGE was particularly superior in classifying whether a data zone would have house prices above or below the mean.

In the realm of explainability, “Comparative Illness Factor” and “Detached Dwellings” were found as the two most important factors for the decisions of the model according to the global explainability results. Specifically, “Comparative Illness Factor” negatively affects house prices while the proportion of “Detached Dwellings” is the most important feature positively influencing house prices. These findings are consistent with our previous work that used XGBoost and XAI to predict and explain house prices [18]. In addition, the surrogate model also highlighted the effect of various types of dwellings, such as the proportion of “Flats” and “Occupied Households” in the prediction of the house prices. However, local explainability results indicated that the factors that are globally important may be less influential locally, in our case, in specific data zones. In the example provided in this work, two nearby data zones, both with average house prices expected to be higher than Scotland’s mean, were influenced by different factors compared to each other and the factors predicted by the global explainability. This underscores the necessity for policymakers to go beyond global explainability and consider the local, geographically specific characteristics and factors relevant to the area of interest when making real-life decisions. In addition, local explainability showed that house prices in a data zone could be influenced by the characteristics and factors of neighboring data zones, indicating interactions between geographically connected areas. Therefore, addressing local issues should not be carried out in isolation by different local administrations. Instead, a holistic approach is required where neighboring public administrations collaborate to carefully select and address factors that are proxies for social issues in their administrations (e.g., poverty, lack of education). This cooperative strategy is essential not only for the house price predictions discussed in this work but also for tackling broader social issues faced by public administrations.

While the similarities between the findings of the GNNExplainer and the SHAP-based approach were substantial, especially with regard to top contributors to feature importance, there were also distinct differences. These differences can be traced back to the intrinsic variations in the methods used in the two models. Specifically, the XGBoost-SHAP approach works with tabular data, treating each data zone as an isolated entity, whereas the GNN-GNNExplainer framework utilizes the underlying graph structure of the data, allowing for the capture of more intricate dependencies between data zones. This distinction is made clear in the context of local explainability, where GNNExplainer pinpointed different features as most important compared to the XGBoost-SHAP analysis. Moreover, the GNNExplainer provides additional information in the form of explanation graphs for each target node. This is a significant advantage over the SHAP approach, which lacks the topological context. However, the difference between the two is not a mere matter of one being superior to the other; rather, it emphasizes the unique insights that can be garnered when considering or not the topology and complex dependencies within data zones. Moreover, the fact that both explainability frameworks produced similar results for top contributors in the predictions is indicative of the robustness of these top contributors across different machine learning models, such as “Comparative Illness factor” and “Detached Dwellings”. However, it also illustrates that the choice of the machine learning model and explainability framework can yield varying insights, especially on a local level, thereby emphasizing the importance of a comprehensive approach when performing such analyses.

Regarding prediction accuracy, it is noteworthy that the GNN variants achieved a marginally higher classification performance than the XGBoost model, with GraphSAGE outperforming XGBoost by a relatively slim margin of 3.6%. While this result might not be substantial, it underscores the potential value of modeling geographical regions as graphs.

For future research in house price prediction, the exploration of additional data-zone-specific features could be instrumental in further improving the accuracy and enhancing
the explainability of GNNs for this particular task. In addition, the presence of a hierarchy in the geographical dimension of the dataset could also be exploited to construct a hierarchical graph similarly to [135] and obtain better results. This study therefore serves as an important stepping stone, highlighting the promise of GNNs and their explainability frameworks in the context of spatial-based prediction tasks.

6. Conclusions

In our analysis, we observed that spatial dependencies significantly influenced house price predictions, indicating that adjacent data zones have a notable impact on the prediction outcomes. Specifically, the results from this study provide compelling evidence of the efficacy of graph neural networks in predicting house prices. Moreover, the research delves into both global and local explainability. Toward this end, three GNN variants (GCN, ChebNet, and GraphSAGE) are employed. The comparative analysis of the results reveals that the three GNN variants outperform the conventional models XGBoost and MLP. These findings underline the significance of incorporating the underlying graph structure of Scottish data zones, with GraphSAGE demonstrating the highest accuracy, precision, recall, and F1 score among all models. GraphSAGE’s early high accuracy scores indicate a robust learning mechanism that efficiently harnesses the local neighborhood information of each node, thereby reducing the model’s training time. The global explainability analysis suggests that features related to living and health conditions, such as “Comparative Illness Factor”, “School attendance”, “Mothers never smoked”, and “Mothers currently smoking”, and housing types such as “Detached dwellings”, “Flats”, “Occupied households”, and “Terraced households”, play a significant role in determining house prices. The local explainability of the model showed the influential nodes and features contributing to the prediction. Interestingly, the local explanations align well with the global interpretations, suggesting consistency in the model’s decision-making process. However, there are some cases where global contributors are less important in specific data zones.

In conclusion, this research contributes valuable insights into the use of graph neural networks for house price prediction. Our findings support the potential of these methods to not only deliver superior prediction performance but also to provide a clear understanding of the factors driving these predictions. While some indicators, such as “Mothers have never smoked” or “Mothers currently smoke”, may not intuitively seem relevant to the prediction outcome, their statistical significance in our model suggests potential indirect relationships or correlations that warrant further investigation. At the same time, in real-life decision making, not all statistical indicators hold the same importance for policy makers, and many do not lend themselves to policy intervention (e.g., “Mothers have never smoked”). This highlights the need for a rigorous selection process for statistical indicators, ensuring that each chosen indicator is justified not only by statistical significance but also by its theoretical relevance to housing market dynamics. Future research should focus on statistical indicators that serve as proxies for important social issues faced by public administrations and can be influenced through policy interventions. In addition, the selected indicators should be shown to have significant impacts (either positive or negative) on predicting house prices based on the results of both global and local explainability analyses. Future work should also focus on enhancing the explainability of GNN models and develop more scalable solutions. Moreover, future work can further investigate the use of GNNs in other prediction tasks, such as node regression, and explore additional strategies for enhancing their explainability. Specifically, we aim to extend our analysis to spatio-temporal regression models, which will enable us to utilize traditional regression metrics and probabilistic frameworks to improve the accuracy and robustness of house price predictions over time.
Author Contributions: E.K. and A.K.: Conceptualization, methodology, writing—review and editing; A.K. and P.B.: software, data curation, writing—original draft; E.K. and K.T.: Project administration, supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “2nd Call for H.F.R.I. Research Projects to support Faculty Members & Researchers” (Project Number: 2412).

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The study was conducted utilizing data that have been previously published in Karamanou, A.; Kalampokis, E.; Tarabanis, K. Integrated statistical indicators from Scottish linked open government data. Data in Brief 2023, 46, 108779. https://doi.org/10.1016/j.dib.2022.108779, accessed on 30 March 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. The statistical indicators used in the open statistics graph classified by theme.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Statistical Indicator</th>
<th>Details/ Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access to Services</td>
<td>Travel times to GP surgeries by public transport</td>
<td>Minutes/Numeric</td>
</tr>
<tr>
<td></td>
<td>Travel times to post office by public transport</td>
<td>Minutes/Numeric</td>
</tr>
<tr>
<td></td>
<td>Travel times to retail centre by public transport</td>
<td>Minutes/Numeric</td>
</tr>
<tr>
<td></td>
<td>Travel times to petrol station by car</td>
<td>Minutes/Numeric</td>
</tr>
<tr>
<td></td>
<td>Travel times to post office by car</td>
<td>Minutes/Numeric</td>
</tr>
<tr>
<td></td>
<td>Travel times to GP surgeries by car</td>
<td>Minutes/Numeric</td>
</tr>
<tr>
<td></td>
<td>Travel times to primary school by car</td>
<td>Minutes/Numeric</td>
</tr>
<tr>
<td></td>
<td>Travel times to secondary school by car</td>
<td>Minutes/Numeric</td>
</tr>
<tr>
<td></td>
<td>Travel times to retail centre by car</td>
<td>Minutes/Numeric</td>
</tr>
<tr>
<td>Crime and Justice</td>
<td>Chimney fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Dwelling fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Other building fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Other primary fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Outdoor fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Refuse fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Vehicle fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Accidental chimney fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Accidental dwelling fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Accidental other building fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Accidental other primary fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Accidental outdoor fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Accidental refuse fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Accidental vehicle fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Not accidental chimney fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Not accidental dwelling fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Not accidental other building fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Not accidental other primary fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Not accidental outdoor fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Not accidental refuse fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Not accidental vehicle fires</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Crime indicators</td>
<td>Ratio/Numeric</td>
</tr>
</tbody>
</table>
Table A1. Cont.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Statistical Indicator</th>
<th>Details/ Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic Activity, Benefits, and Tax Credits</td>
<td>Children 0–15 living in low-income families</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Children 0–19 living in low-income families</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Age of first-time mothers 19 years and under</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Age of first-time mothers 35 years and older</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Employment deprivation</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td>Education, Skills, and Training</td>
<td>School attendance</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Educational attainment of school leavers</td>
<td>Score/Numeric</td>
</tr>
<tr>
<td>Geography</td>
<td>Land area</td>
<td>Hectares/Numeric</td>
</tr>
<tr>
<td></td>
<td>Urban rural classification</td>
<td>6-fold/Categorical</td>
</tr>
<tr>
<td>Health and Social Care</td>
<td>Mothers currently smoking</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Mothers former smokers</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Mothers never smoked</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Low birth-weight (less than 2500 g) babies (single births)</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Not known if mothers smoked</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Comparative illness factor</td>
<td>~/Integer</td>
</tr>
<tr>
<td>Housing</td>
<td>Dwellings per hectare</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Detached dwellings</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Flats</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Semi-detached dwellings</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Terraced dwellings</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Dwellings of unknown type</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Long-term empty households</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Occupied households</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Second-home households</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Vacant households</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Households with occupied exemptions</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Households with unoccupied exemptions</td>
<td>Ratio/Numeric</td>
</tr>
<tr>
<td></td>
<td>Households with single adult discounts</td>
<td>Ratio/Numeric</td>
</tr>
</tbody>
</table>

Table A2. Paired t-test results for performance metrics across different models.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Metric</th>
<th>t-Statistic</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GraphSAGE vs. GCN</td>
<td>Accuracy</td>
<td>10.203722</td>
<td>4.168358 × 10^{-11}</td>
</tr>
<tr>
<td>GraphSAGE vs. GCN</td>
<td>Precision</td>
<td>9.183735</td>
<td>4.391700 × 10^{-10}</td>
</tr>
<tr>
<td>GraphSAGE vs. GCN</td>
<td>Recall</td>
<td>8.531244</td>
<td>2.127801 × 10^{-9}</td>
</tr>
<tr>
<td>GraphSAGE vs. GCN</td>
<td>F1</td>
<td>9.622985</td>
<td>1.566780 × 10^{-10}</td>
</tr>
<tr>
<td>GraphSAGE vs. GCN</td>
<td>ROC-AUC</td>
<td>15.862256</td>
<td>7.876512 × 10^{-16}</td>
</tr>
<tr>
<td>GraphSAGE vs. ChebNET</td>
<td>Accuracy</td>
<td>15.184385</td>
<td>2.450008 × 10^{-15}</td>
</tr>
<tr>
<td>GraphSAGE vs. ChebNET</td>
<td>Precision</td>
<td>10.012552</td>
<td>6.41638 × 10^{-11}</td>
</tr>
<tr>
<td>GraphSAGE vs. ChebNET</td>
<td>Recall</td>
<td>12.661661</td>
<td>2.432774 × 10^{-13}</td>
</tr>
<tr>
<td>GraphSAGE vs. ChebNET</td>
<td>F1</td>
<td>9.778460</td>
<td>1.094454 × 10^{-10}</td>
</tr>
<tr>
<td>GraphSAGE vs. ChebNET</td>
<td>ROC-AUC</td>
<td>9.956104</td>
<td>7.291891 × 10^{-11}</td>
</tr>
<tr>
<td>GraphSAGE vs. XGBoost</td>
<td>Accuracy</td>
<td>14.172647</td>
<td>1.437887 × 10^{-14}</td>
</tr>
<tr>
<td>GraphSAGE vs. XGBoost</td>
<td>Precision</td>
<td>9.746166</td>
<td>1.178822 × 10^{-10}</td>
</tr>
<tr>
<td>GraphSAGE vs. XGBoost</td>
<td>Recall</td>
<td>18.184365</td>
<td>2.128981 × 10^{-17}</td>
</tr>
<tr>
<td>GraphSAGE vs. XGBoost</td>
<td>F1</td>
<td>11.667883</td>
<td>1.786180 × 10^{-12}</td>
</tr>
<tr>
<td>GraphSAGE vs. XGBoost</td>
<td>ROC-AUC</td>
<td>7.186115</td>
<td>6.546190 × 10^{-8}</td>
</tr>
<tr>
<td>GraphSAGE vs. MLP</td>
<td>Accuracy</td>
<td>20.814211</td>
<td>5.551926 × 10^{-19}</td>
</tr>
<tr>
<td>GraphSAGE vs. MLP</td>
<td>Precision</td>
<td>18.070265</td>
<td>2.518977 × 10^{-17}</td>
</tr>
<tr>
<td>GraphSAGE vs. MLP</td>
<td>Recall</td>
<td>21.193442</td>
<td>3.393254 × 10^{-19}</td>
</tr>
<tr>
<td>GraphSAGE vs. MLP</td>
<td>F1</td>
<td>19.587716</td>
<td>2.885125 × 10^{-18}</td>
</tr>
<tr>
<td>GraphSAGE vs. MLP</td>
<td>ROC-AUC</td>
<td>8.411927</td>
<td>2.856781 × 10^{-9}</td>
</tr>
</tbody>
</table>
Table A2. Cont.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Metric</th>
<th>t-Statistic</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCN vs. ChebNET</td>
<td>Accuracy</td>
<td>2.949306</td>
<td>6.238148×10^{-3}</td>
</tr>
<tr>
<td>GCN vs. ChebNET</td>
<td>Precision</td>
<td>1.905649</td>
<td>6.665262×10^{-2}</td>
</tr>
<tr>
<td>GCN vs. ChebNET</td>
<td>Recall</td>
<td>3.255215</td>
<td>2.880711×10^{-3}</td>
</tr>
<tr>
<td>GCN vs. ChebNET</td>
<td>F1</td>
<td>0.065378</td>
<td>9.483214×10^{-1}</td>
</tr>
<tr>
<td>GCN vs. ChebNET</td>
<td>ROC-AUC</td>
<td>-6.174785</td>
<td>9.871698×10^{-7}</td>
</tr>
<tr>
<td>GCN vs. XGBoost</td>
<td>Accuracy</td>
<td>4.771361</td>
<td>4.787868×10^{-5}</td>
</tr>
<tr>
<td>GCN vs. XGBoost</td>
<td>Precision</td>
<td>1.649143</td>
<td>1.099108×10^{-1}</td>
</tr>
<tr>
<td>GCN vs. XGBoost</td>
<td>Recall</td>
<td>7.258198</td>
<td>5.417471×10^{-8}</td>
</tr>
<tr>
<td>GCN vs. XGBoost</td>
<td>F1</td>
<td>3.736712</td>
<td>8.136250×10^{-4}</td>
</tr>
<tr>
<td>GCN vs. XGBoost</td>
<td>ROC-AUC</td>
<td>-9.460600</td>
<td>2.286692×10^{-10}</td>
</tr>
<tr>
<td>GCN vs. MLP</td>
<td>Accuracy</td>
<td>9.801749</td>
<td>1.037470×10^{-10}</td>
</tr>
<tr>
<td>GCN vs. MLP</td>
<td>Precision</td>
<td>10.488417</td>
<td>2.212724×10^{-11}</td>
</tr>
<tr>
<td>GCN vs. MLP</td>
<td>Recall</td>
<td>9.452620</td>
<td>2.329783×10^{-10}</td>
</tr>
<tr>
<td>GCN vs. MLP</td>
<td>F1</td>
<td>12.171377</td>
<td>6.415196×10^{-13}</td>
</tr>
<tr>
<td>GCN vs. MLP</td>
<td>ROC-AUC</td>
<td>-7.182763</td>
<td>6.604146×10^{-8}</td>
</tr>
<tr>
<td>ChebNET vs. XGBoost</td>
<td>Accuracy</td>
<td>3.137682</td>
<td>3.888951×10^{-3}</td>
</tr>
<tr>
<td>ChebNET vs. XGBoost</td>
<td>Precision</td>
<td>-0.121202</td>
<td>9.043668×10^{-1}</td>
</tr>
<tr>
<td>ChebNET vs. XGBoost</td>
<td>Recall</td>
<td>3.416474</td>
<td>1.897231×10^{-3}</td>
</tr>
<tr>
<td>ChebNET vs. XGBoost</td>
<td>F1</td>
<td>3.960737</td>
<td>4.450284×10^{-4}</td>
</tr>
<tr>
<td>ChebNET vs. XGBoost</td>
<td>ROC-AUC</td>
<td>-3.098158</td>
<td>4.298137×10^{-3}</td>
</tr>
<tr>
<td>ChebNET vs. MLP</td>
<td>Accuracy</td>
<td>8.517380</td>
<td>2.201690×10^{-9}</td>
</tr>
<tr>
<td>ChebNET vs. MLP</td>
<td>Precision</td>
<td>6.984752</td>
<td>1.114174×10^{-7}</td>
</tr>
<tr>
<td>ChebNET vs. MLP</td>
<td>Recall</td>
<td>7.597748</td>
<td>2.239849×10^{-8}</td>
</tr>
<tr>
<td>ChebNET vs. MLP</td>
<td>F1</td>
<td>11.477786</td>
<td>2.649459×10^{-12}</td>
</tr>
<tr>
<td>ChebNET vs. MLP</td>
<td>ROC-AUC</td>
<td>-1.054002</td>
<td>3.005850×10^{-1}</td>
</tr>
<tr>
<td>XGBoost vs. MLP</td>
<td>Accuracy</td>
<td>5.436311</td>
<td>7.548705×10^{-6}</td>
</tr>
<tr>
<td>XGBoost vs. MLP</td>
<td>Precision</td>
<td>6.991994</td>
<td>1.092980×10^{-7}</td>
</tr>
<tr>
<td>XGBoost vs. MLP</td>
<td>Recall</td>
<td>3.976198</td>
<td>4.267591×10^{-4}</td>
</tr>
<tr>
<td>XGBoost vs. MLP</td>
<td>F1</td>
<td>6.532380</td>
<td>3.738756×10^{-7}</td>
</tr>
<tr>
<td>XGBoost vs. MLP</td>
<td>ROC-AUC</td>
<td>2.148373</td>
<td>4.016817×10^{-2}</td>
</tr>
</tbody>
</table>

References


6. Pereira, P.; Zhao, W.; Symochko, L.; Inacio, M.; Bogunovic, I.; Barcelo, D. The Russian-Ukrainian armed conflict will push back the sustainable development goals. *Geogr. Sustain.* 2022, 3, 277–287. [CrossRef]


22. Taecharungroj, V. Google Maps amenities and condominium prices: Investigating the effects and relationships using machine learning. *Habitat Int.* 2021, 118, 102463. [CrossRef]


43. Brimos, P.; Karamanou, A.; Kalampokis, E.; Tarabanis, K. Graph Neural Networks and Open-Government Data to Forecast Traffic Flow. *Information* 2023, 14, 228. [CrossRef]


131. Karamanou, A.; Kalampokis, E.; Tarabanis, K. Integrated statistical indicators from Scottish linked open government data. Data Brief 2023, 46, 108779. [CrossRef]


133. Liu, X. Spatial and Temporal Dependence in House Price Prediction. J. Real Estate Financ. Econ. 2013, 47, 341–369. [CrossRef]


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.