Exploring Public Transport Trip Attraction factors across Greater London using Interpretable Machine Learning on Open Data

The-Huan Hoang
Supervised by Dr. Elsa Arcaute

A dissertation presented in partial fulfillment
of the requirements for the degree of
Master of Science in Urban Spatial Science

Bartlett Centre for Advanced Spatial Analytics (CASA)
University College London
United Kingdom
23 August 2024

Abstract

Abstract goes here...

Declaration

I, The-Huan Hoang, declare that the work presented in this dissertation is my own. Where claims, information or findings have been derived from other authors and sources, I confirm that this has herein been indicated in writing, attributing due credit. In the spirit of fostering reproducible research, all scripts used for analysis and their outputs are made available on GitHub at https://github.com/shaunhoang/casa-tfl-dissertation.

Acknowledgements

I am grateful for the guidance and insightful feedback from my supervisor, Dr. Elsa Arcaute, who has shaped the direction of my work from its inception. I would also like to acknowledge the invaluable support from the Public Transport Service Planning team at Transport for London. Their expert advice and real-world insights into the data science behind transport planning have greatly enriched my knowledge, as well as this research project.

Contents

1 Introduction **8**

2 Literature Review **9**
2.1 Urban amenities as attractors . . . 10
2.2 Transport accessibility as attractors . . . 11
2.3 Public transport and the walkable city . . . 11
2.4 Explainable Artificial Intelligence for spatial systems . . . 12
2.5 Open data and reproducibility . . . 13
2.6 Summary . . . 14

3 Methodology **15**
3.1 Defining the spatial unit of analysis . . . 16
3.2 Data Preprocessing . . . 18
3.2.1 Aggregating target variables . . . 18
3.2.2 Feature engineering . . . 19
3.2.3 Incorporating spatial lags as a feature . . . 23
3.3 Model selection and interpretation . . . 26

4 Results **29**

5 Discussion **38**

6 Conclusion **41**

A Appendix **42**
A.1 Open data sources . . . 42
A.2 True to Data and True to Model in SHAP technique . . . 44
A.3 Supplemental Figures . . . 45

List of Figures

3.1 Overview of the feature engineering pipeline 15
3.2 Overview of model training and interpretation 16
3.3 Methodology to generate 15-minute walking isochrones as the spatial unit of analysis 17
3.4 Two example spatial units of analysis located in Central London 17
3.5 Public transport arrivals on a typical Saturday, aggregated to chosen spatial units of analysis. *Top* Total, *Bottom* by time band. 18
3.6 Distribution of the Total arrivals as raw count (left) and log-transformed (right). Similar transformations are carried out for arrivals by each of the four time bands 19
3.7 Population density (UK Census 2011) 20
3.8 Amenity and Transport POI density by type (Source: OSM) 20
3.9 Normalised diversity indices for Amenity POIs, Transport POIs by mode and by service. *Top*: Mapping top 10% values, *Bottom*: Distribution of diversity indices . 22
3.10 Top 10% values of three diversity indices Borough: Hammersmith and Fulham . . 22
3.11 Network graphs of TfL Rail (left) and Bus (right) networks (Source: TfL) 23
3.12 Mapping centrality measures of spatial units (isochrones) on Rail (left) and Bus (right) networks 24
3.13 LISA cluster map of Total arrivals showing local sptial autocorrelation Global Moran's I: 0.742 (High spatial autocorrelation) 25
3.14 Spatial vs. non-spatial nested cross-validation using 5-folds for hyperparameter tuning and performance evaluation. This technique helps ensure the training and test sets in both the inner and the outer folds are spatially clustered to avoid data leakage. Source: Schratz et al. (2018) 27

4.1 Correlation matrix of the input features and target variable Total arrivals (log) . . 30
4.2 Observations with significant overprediction and underprediction (N=576 of 16,890) Fine-tuned XGBoost model, Total arrivals (log) 31
4.3 Summary of feature importance based on SHAP values of a sample. Red-Blue scale represents the feature value, and the x-axis represents the SHAP value, i.e., impact on final model prediction of Total arrivals (log) 32
4.4 Feature importance in predicting Total arrivals (log) Features clustered by correlation and ordered by mean absolute SHAP values 33
4.5 Local SHAP values for 2 spatial units Kew Gardens and Covent Garden 34
4.6 Mapping local SHAP values for 2 features: Outdoor leisure POIs and Entertainment leisure POIs. Target: Total arrivals (log) 35
4.7 Mapping local SHAP values for 2 features: Outdoor leisure POIs and Entertainment leisure POIs, weighted by number of arrivals 36
4.8 Raw vs. weighted SHAP for 2 features Outdoor leisure POIs and Entertainment leisure POIs Total arrivals (log) 37

5.1 Hot spots for destination groups: Amenity vs. Residential vs. Transport Midday vs Late . . . 38
5.2 Identifying high activity areas in Wood Green . . . 39

A.1 Map of the spatial k-fold cross-validation splits for hyperparameter tuning and performance assessment (k=10) . . . 45
A.2 Summary of feature importance based on SHAP values of a sample. Red-Blue scale represents the feature value, and the x-axis represents the SHAP value, i.e., impact on final model prediction of arrivals in each time band (log) . . . 45

List of Tables

3.1 Summary of the 25 main features engineered for the model 24
3.2 Mapping spatial lag features to predict corresponding target variables 26

4.1 Comparing model R^2 results of 3 models predicting Total arrivals (raw) and Total arrivals (log) using spatial k-fold cross-validation 29
4.2 Finetuned XGBoost model evaluation results for All target variables (log) using nested spatial k-fold cross-validation 30

A.1 Summary of data used in the analysis and their sources 42
A.2 Summary of Amenity and Transport POI types used in the analysis 44

1. Introduction

2. Literature Review

Estimating trip attraction is a crucial topic of the transport planning academic literature and profession, constituting an integral component of the demand forecast framework, termed the Four-Step Model (FSM), which has been in ubiquitous use since the 1950s (*Travel Forecasting Resource* 2024). More specifically, the FSM begins with the Trip Generation step, where the number of trips generated by each traffic analysis zone is estimated, and Trip Distribution allocates these trips to their destination zones before proceeding to mode choice determination and route assignment. Planners accomplish the task of Trip Distribution using a variety of spatial interaction models, most prominent among which is the classic *Gravity Model* and its derivatives inspired by Newton's law of universal gravitation (Erlander and Stewart, 1990). Originally formulated to estimate flows among cities in a region, it has been adapted to intraurban mobility and suggests that the number of trips between two locations is proportional to the difference in *mass* of each location and inversely proportional to the distance between them.

However, it is worth noting that the definition of *mass* in any gravity-based spatial interaction model has seen many interpretations and experimentations, depending on the types of travel demand of research interest. For example, a spatial interaction model for commuting trip forecasts may consider the emissivity of residential areas as origins and the attractiveness of employment centres as destinations. In fact, the estimation of trip attraction in transport planning literature has predominantly revolved around commutes (i.e., primary travel demand) from residential zones to zones where employment centres or education institutions are located. On the one hand, this choice is a practical one first and foremost, as commuting trips represent a large share of regular intracity mobility flows, whose high volume within short peak periods has important implications on road capacity management or public transit operations, among others. Due to its importance, the calibration of the spatial interaction model for commuting flows is further aided by nationwide flow datasets, such as that present in the UK Census Data[1], which provides detailed and explicit Origin-Destination data to serve as ground truth. In the absence of this declared data, researchers have also turned to spatial features to calibrate commute trip distribution, as explored by Yang et al., 2014.

On the other hand, a plethora of potential trips generated for other purposes is often underresearched, partially due to the complexity of non-commute trip purpose identification, classification and quantification: Non-commute trips can be defined as trips that are not carried out to fulfil time-bound and regular work or education obligations and can include trips for shopping, leisure, and social activities. Therefore, as opposed to commuting trips, Non-commute trips are highly irregular both spatially and temporally, i.e., they can vary in terms of distance, duration, and mode of transport from one to another, even for the same destination between different days.

[1] An interactive visualisation of the UK Census 2021 origin-destination data can be found at https://www.ons.gov.uk/visualisations/censusorigindestination/

Nevertheless, Non-commute trips are an important component of urban travel behaviour and can have a significant impact on the overall transportation system. Advances in the field have brought about improvements to methodologies such as travel diary surveys[2], which collect data on personal and household travel behaviour, including Non-commute trips, to allow for comprehensive planning. By design, however, they are limited in scale and generalisability.

Overall, this reveals a gap in our understanding of the comprehensive travel behaviour of city dwellers on an aggregate scale. With the number of trips for shopping, leisure, and social activities increasing in recent years and the number of work-related trips decreasing as flexible work arrangements become more prevalent (Wöhner, 2022), with recognised health benefits to workers (MacLeod, Cole, and Musselwhite, 2022), it is important to understand the factors that influence non-commute trip attraction and to incorporate them into transport planning models. Within the scope of this research, we hypothesise that a given urban area's non-commute trip attracting *mass* is associated with the built environments, which include their amenity and public transport accessibility profiles.

2.1 Urban amenities as attractors

Cervero and Kockelman (1997) was one of the foundational works to provide an econometric estimation of how three groups of factors pertaining to the built environments—namely Density, Diversity and Design—are associated with travel demand to a traffic analysis zone for different purposes (work/non-work) and with different mode choices (private vehicles/public transport). This study is novel for demonstrating that changes in the built environment—such as density, diversity, and walkability have a tangible impact on travel behaviour, which supports the idea that designing more compact, mixed-use, pedestrian-friendly neighbourhoods can significantly influence how people choose to travel. The components of density and diversity refer to the number and the variety of destinations available in a given area, respectively, while the design component refers to the physical layout of the area, including the presence of sidewalks, bike lanes, and other pedestrian-friendly infrastructure. It is worth noting that the study considered the distance as a control impedance factor—adhering to the classic Gravity Model paradigm—and did not investigate how these factors may interact with each other.

There have been many recent attempts to link the presence of urban amenities to trip attraction and calibrate how we define urban attractors. Beyond the pure correlation observed between the density of points of interest (POIs) in a region and intensity of trips made to a given area (Melikov et al., 2021), the typology of the POIs differs between urban clusters that see distinct patterns of inflow trips made. For example, after Aaqib Javed et al. (2020) have clustered urban areas in one city into Global (e.g., airports), Downtown (d.g., central business district) and Residential attractors based on the volume, spatial dispersion and distance of the trips made using O-D data from mobile phone, they found that the differences in the types of POIs in each of these clusters were also statistically significant. The typology of amenity POIs explored in this study included not only Retail, as is common when thinking about non-commute trips, but also hospitals, public services, restaurants, religious institutions, and other categories. This suggests that the type of urban amenities available in an area can influence the volume of trips made to that area and that different types of amenities may attract different types of trip purposes.

Lastly, we must also consider the femininity diversity factor, apart from the typology of the

[2]An example of these is the UK National Travel Survey, accessible at https://www.gov.uk/government/collections/national-travel-survey-statistics

amenities themselves, in the amenity profile of areas as a potential attractor, following the conclusion made by Cervero and Kockelman (1997). More specifically, for non-commute trips where locations are not necessarily prescribed, and the destination choice can be influenced by personal preferences and circumstantial convenience, people may be more likely to visit more vibrant areas that offer a variety of services and activities where multiple tasks could be carried out, as opposed to areas where one type of destinations is overrepresented. This subjective perception of the quality of urban areas is consistent with the concept of 15-minute cities, which purports that mixed-use areas with a diversity of amenities within a 15-minute walking distance are more attractive areas to live in and, by extension, to visit than single-use and exclusive ones. (Khavarian-Garmsir et al., 2023)

2.2 Transport accessibility as attractors

If the density and diversity of amenities were the only explanatory factors for non-commute trip demand, it would be tempting to conclude that the local configuration of destination points of interest is the sole determinant of why people go to one place and not another for their diverse needs. In reality, neighbourhoods do not exist in a vacuum but are intricately connected to one another by the transportation network. In simple terms, the consideration of distance as friction in the classic Gravity Model attests to the fact that the more accessible a destination is, the more likely it is to attract travel demand, especially trips with higher elasticity of choice such as non-commute (i.e., trips made to a specific destination not based on obligation, but mainly out of convenience). Public transport accessibility has been shown to contribute to the growth of intercity importance of certain areas as opposed to others and preserve travel demand variability over time (Zhong et al., 2015), which means that it reinforces the prominence of well-connected areas as urban attractors, presumably with similar density and diversity of destination points of interest.

Jayasinghe, Sano, and Rattanaporn (2017) went one step further to purport that the network centrality property of an urban area is enough to predict travel demand to a high degree of accuracy without the need to consider other attractors such as employment centres and amenities, thanks to its tight correlation. More specifically, the study looked at four centrality measures: Connectivity, Choice, Global Integration and Local Integration, which are analogous to the concept of degree centrality, betweenness centrality, closeness centrality, and closeness centrality in a confined radius, respectively. Among them, Global Integration (closeness centrality) was found to be the most significant predictor, suggesting that the role of the transport network in shaping urban travel demand at the aggregate scale is paramount, perhaps just as important as socio-economic factors (Convery and Williams, 2019)

2.3 Public transport and the walkable city

Walkability contributes to the quality of life in cities. The 15-minute city is one of the most recent urban planning concepts that has gained traction in recent years. The idea is to create more sustainable and resilient cities by reducing the need for long commutes and promoting active modes of transport such as walking and cycling. A resident who lives in a neighbourhood that has all the amenities they need within a 15-minute walk or cycle can ideally reduce their reliance on cars and public transport to a minimum. However, this does not mean that the role of public transport is diminished within the 15-minute city vision. On the contrary, public transport is an essential component of a sustainable system, which provides an efficient and affordable way

for people to travel longer distances, effectively agglomerating a larger urban system made up of ideally 15-minute cities. In other words, public transport is the backbone of the 15-minute city, connecting neighbourhoods and enabling people to travel further than they can on foot or by bike. This, in turn, has multiple implications for the local communities:

- From the lens of urban planning and traffic management, the higher the volume of public transport trips made to an area, the higher the amount of pedestrian traffic it receives on top of the local population who patronises their local amenities. Understanding the trip attraction factors can help prioritise investments in public transport capacity planning, pedestrian or cycling infrastructures and services for these individuals, which can lead to a virtuous cycle of increased active travel and reduced car dependency.
- From the lens of fomenting local economies, increased pedestrian traffic injected by public transport can also be seen as an additional source of footfall, which is invaluable for businesses with storefronts. Findings on the economic benefits of walking and cycling done by Transport for London[3] showed that retail rents increased by 7%, and office space rents increased by 4% in areas that encouraged non-motorised access, suggesting that the street improvements translated into much more desirable spaces.

It is also worth mentioning that footfall estimation itself is a prolific field of research within retail analytics, which counts on a variety of technologies such as WiFi tracking, video analytics, and mobile phone data to estimate the number of people who pass by a given location. The insights derived from footfall data can be used to inform decisions on store location, opening hours, staffing levels, and marketing strategies, among others. Uncovering patterns of public transport trip attraction, which in turn is an indicator of pedestrian influx, could be used to complement footfall data and provide a more comprehensive understanding of the factors that influence the number of people in a given area.

2.4 Explainable Artificial Intelligence for spatial systems

Up to this point, our visited literature has shown that both urban amenities and transport accessibility are important factors in determining the attractiveness of an area to non-commute trips, albeit in isolation. In reality, the interaction between these two factors has not been well studied in conjunction. It is possible that the dense presence of diverse urban amenities is brought about by preexisting favourable transport accessibility or vice versa. In the interest of addressing the research questions, we are presented with a dilemma of interpretability versus prediction accuracy in the choice of modelling techniques.

Spatial statistical methods, commonly used in policy analysis, can help us quantify and interpret the relationship between the dependent variables and independent variables in the form of explicit coefficients. However, even spatially aware methods such as Spatial Lag Model (SLM) or Geographically Weighted Regression (GWR) are limited in their linear assumptions of linearity, issues with multicollinearity, and thus may fail at attaining high prediction accuracy (Wheeler and Tiefelsdorf, 2005).

In the meantime, machine learning methods have gained popularity in the field of urban mobility research due to their ability to capture complex, non-linear relationships between spatial features with high prediction accuracy. However, the lack of interpretability of machine learning

[3]https://tfl.gov.uk/corporate/publications-and-reports/economic-benefits-of-walking-and-cycling

models has been a major drawback, especially in the context of urban planning, where the ability to understand the underlying factors that drive the model's predictions is crucial for decision-making. Among many opportunities to develop Machine Learning explanation techniques for causal inference in urban mobility is the need to evaluate causal effects of input features and derive actionable insights (Xin et al., 2022).

To address the interpretability challenge, local interpretation methods like SHAP (SHapley Additive exPlanations)—a method based on cooperative game theory based on the seminal work by Lundberg and Lee, 2017—to provide detailed attributions of input features to individual predictions made. Its introduction contributed to the growing body of literature on interpretable machine learning with spatial systems for regression with XGBoost (Li, 2022) or classification with Light GBM (Louhichi et al., 2023). One of the most notable recent examples of SHAP for urban mobility in action was the development of the Deep Gravity Model introduced by Simini et al., 2021, which used SHAP to explain the predictions of a deep learning model for spatial interaction and how the model's predictions were influenced by the input spatial features such as land-use mix, points of interest in the origin and destination zones.

As we approach this research with an eye on accurately modelling the real-world phenomenon and explaining the contributing factors, instead of GWR or SLM, we will use a machine learning model to predict non-commute trip attraction in urban areas and then leverage SHAP to explain the predictions in an attempt to quantify the relative importance and contribution of urban amenities and transport accessibility.

2.5 Open data and reproducibility

Many recent advances in the field of urban mobility research have been made possible by the availability of proprietary datasets owned by private companies, such as mobile phone data, GPS data, and social media data. Although the high data granularity has opened doors to new research opportunities and the development of new advanced techniques, its use also raises concerns about data privacy, data ownership, and data sharing. Moreover, it heightens the barrier to entry for independent researchers who would like to replicate the studies for their respective cities or regions. In contrast, open data not only democratises independent and participatory research in the public interest but also breaks down many reproducibility barriers when applying established methodologies to other localities (Yadav et al., 2017).

On one side, we have readily available open datasets maintained by local or federal government agencies in many countries. On the other, we have global platforms like **OpenStreetMap** and **Overture Maps Foundation** that provide a wealth of spatial data for cities around the world. The use of open data to foment reproducibility will be a key focus of this research to ensure that the findings can be replicated in other cities and regions, unburdened by commercial data availability and access barriers. It also enables flexible preprocessing of the amenity or transportation features to model the intracity mobility behaviour that is pertinent to the locality's needs while adhering to the same model training and analysis methodology. These needs might include a different classification of the amenity points of interest that correspond to different trip purposes or a different categorisation of the services in the public transport network.

Lastly, adhering to the spirit of openness and future reproducibility across different urban contexts, all of the datasets used in this research will be entirely sourced from open data hubs and platforms. The associated scripts used for data processing, model training and prediction analytics are available on GitHub.

2.6 Summary

In order to address the research questions that we have posed, informed by the literature review thus far, we intend to tune, evaluate and select a suitable machine learning model to predict non-commute trip attraction in urban areas, where we hypothesise that the built environment, including urban amenities and transport accessibility, are important factors. The use of a machine learning model to represent these relationships, in conjunction with explainable machine learning techniques such as SHAP, allows us to capture complex, non-linear relationships between features and achieve high prediction accuracy of the target variable, as well as analyse attributions of input features to the predictions made in order to uncover underlying contributions.

3. Methodology

In this section, we will first describe the methodology used to preprocess data from various sources to engineer the necessary features for the predictive model. Second, we will outline the model selection and evaluation process, and finally, we will discuss the methodology used to interpret the model predictions using SHAP techniques. For ease of reference, the data preprocessing and feature engineering pipeline is summarised in Figure 3.1, and an overview of the model training and prediction interpretation methodology can be found in Figure 3.2. More detailed information on the data sources and the feature engineering process can be found in the Data Sources section of the Appendix.

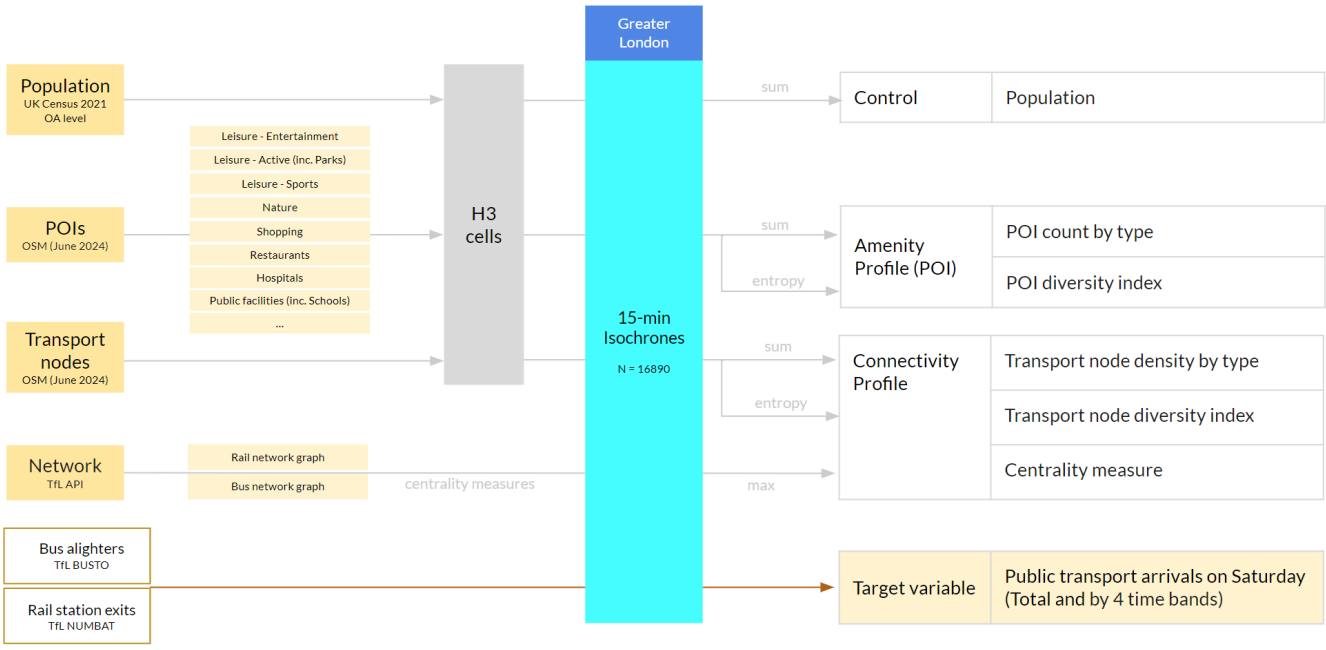

Figure 3.1: Overview of the feature engineering pipeline

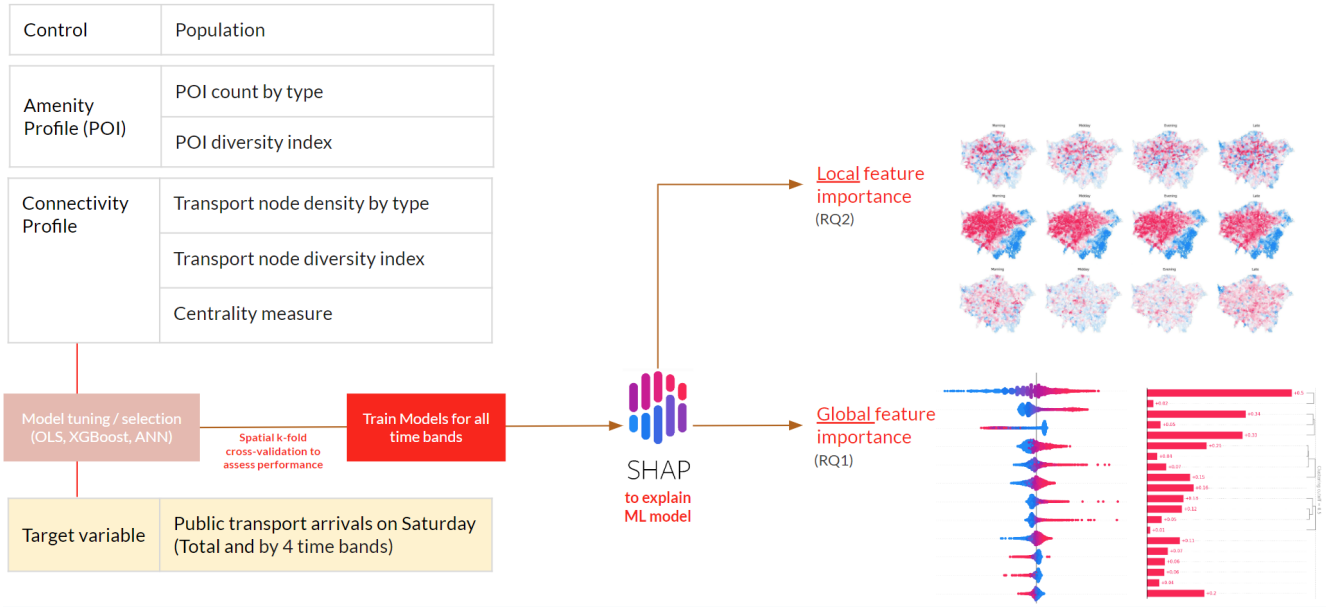

Figure 3.2: Overview of model training and interpretation

3.1 Defining the spatial unit of analysis

As features will be extracted and engineered from different data sources, the spatial unit is an important consideration, at which the features are aggregated and the Machine Learning model is trained. In this analysis, we will use 15-minute walking isochrones—i.e., the area that can be reached within a 15-minute walk from a given point—as the spatial unit of analysis. The isochrones were generated based on initial seeds of 16,890 points across the Greater London area, using the pedestrian profile of the OpenStreetMap Network API (osmnx), which takes into account pedestrian infrastructure such as footpaths, pedestrian crossings, and pedestrianised areas. The choice of using isochrones as the spatial unit of analysis for our model is motivated by the following reasons:

- First, the spatial unit of aggregation and analysis for the model must reflect how individuals interact with transport nodes and amenity POIs that are accessible using the existing pedestrian network. Proximity using the Euclidean distance is inaccurate when there are natural barriers, such as rivers or private estates, that prevent thorough access.
- Second, human mobility in cities is not bound by administrative boundaries such as Output Areas or boroughs, and the use of overlapping isochrones allows us to capture the spatial heterogeneity of the urban environment in our dataset.
- Third, the 15-minute specification of the isochrones aims to reflect a uniformity among the spatial units of analysis while following a predetermined standard set out by the 15-minute city concept that characterises the extent of desirable mobility in a neighbourhood.
- Lastly, the density of the isochrones produced is a compromise between the granularity of the spatial unit and the computational resources required to process the data. The initial seeds of 16,890 points are originally centroids of the same number of H3 cells at resolution 9 needed to cover the Greater London area.

The detailed methodology to generate the isochrones is illustrated in Figure 3.3, and two example isochrones in Central London can be found in Figure 3.4. It is worth noting that depending

on where the spatial unit is located in the Greater London area, the isochrones may vary in shape and size due to the underlying pedestrian network and the distribution of transport nodes and amenity POIs. However, these variances are expected and even desirable as they more closely reflect pedestrian accessibility to different types of amenities and transport nodes. All raw data extracted from various sources will be aggregated and processed at the isochrone level to generate the necessary features and target variable for the model training (See Figure 3.1). In the next section, we will describe the data transformation and feature engineering process in more detail. Note that the terms spatial units (of analysis), isochrones, or simply areas may be used interchangeably from this point on in our discussion.

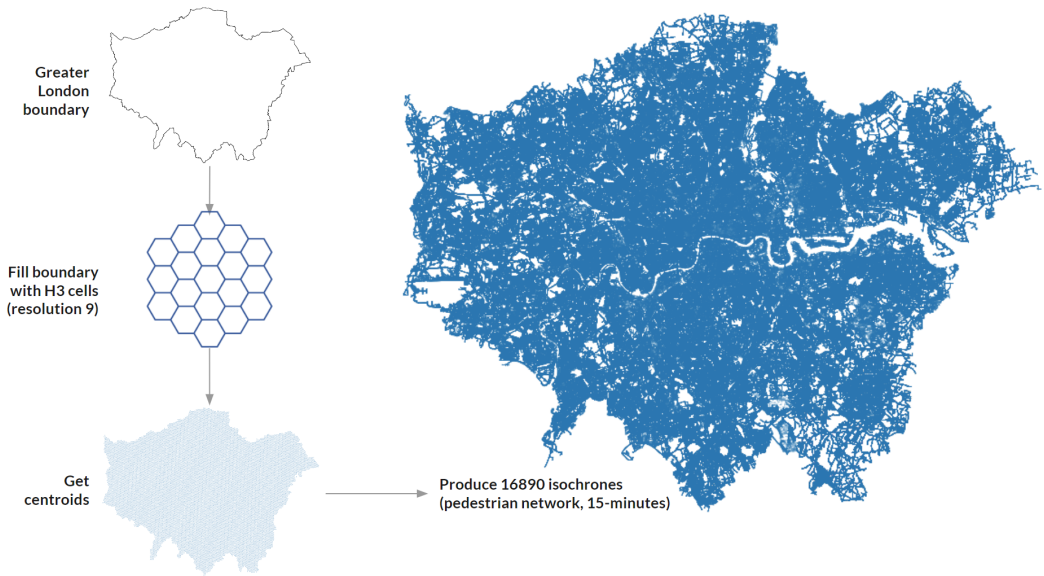

Figure 3.3: Methodology to generate 15-minute walking isochrones as the spatial unit of analysis

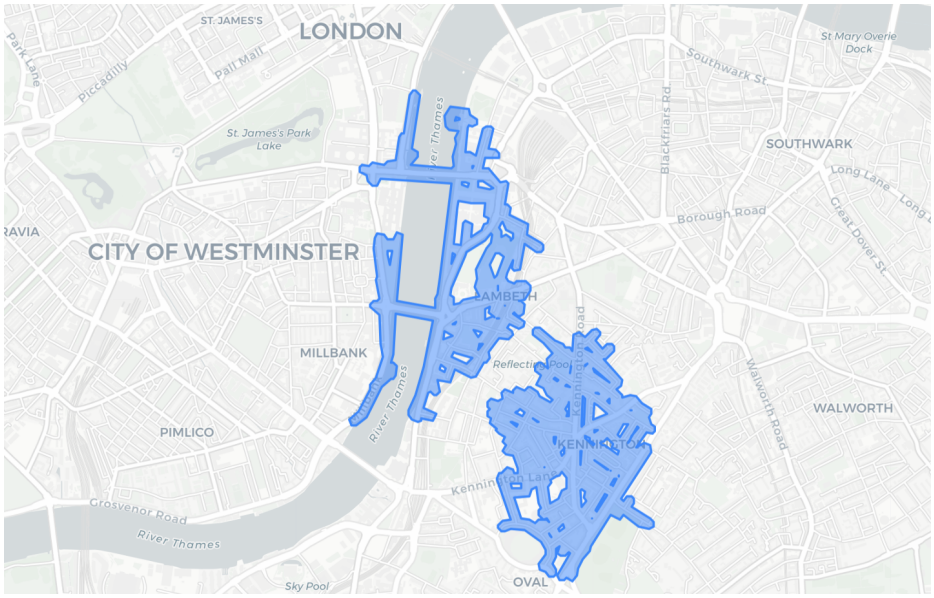

Figure 3.4: Two example spatial units of analysis located in Central London

3.2 Data Preprocessing

3.2.1 Aggregating target variables

The **total arrivals**, i.e., the numbers of passengers who arrive at a transport node or alight from a bus at a stop in a spatial unit (isochrone), are the target variables for our model. The aggregation results are illustrated in Figure 3.5, which shows the total daily figures as well as the figures aggregated into the four time bands described earlier in order to capture the temporal variations. The total daily figures show a clear concentration of passenger arrivals in the Central London area, with the highest number of passengers arriving at transport nodes. The time band figures show a similar pattern, with the highest number of passengers arriving during the Midday time band (10:00-19:00). The Evening time band also shows a significant number of passengers arriving at transport nodes, while the Morning and Late time bands show lower numbers of passengers using the system, due to early hours in the day on weekends and sparser night services, respectively.

Lastly, since the target variable has a highly right-skewed distribution, we will apply a log transformation to the target variable to normalise the distribution and improve the model's performance, which will be addressed in the Results section. The log transformation of the target variable is shown in Figure 3.6, which shows a more normal distribution of the target variable after the transformation and will be used in the model training and evaluation steps implicitly from this point on.

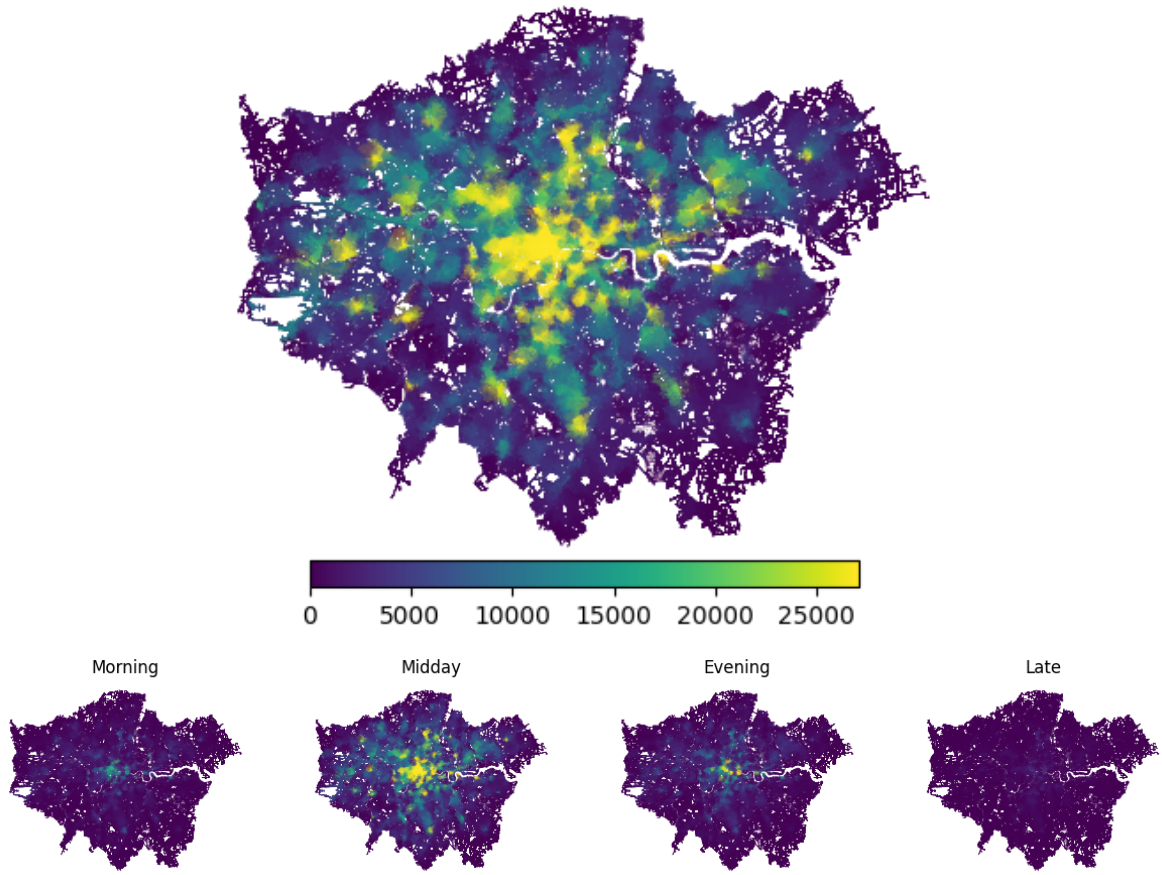

Figure 3.5: Public transport arrivals on a typical Saturday, aggregated to chosen spatial units of analysis. *Top* Total, *Bottom* by time band.

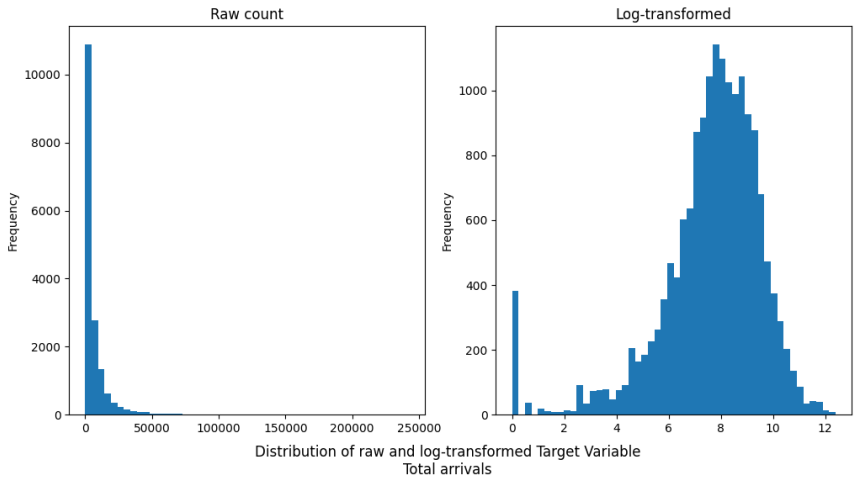

Figure 3.6: Distribution of the Total arrivals as raw count (left) and log-transformed (right). Similar transformations are carried out for arrivals by each of the four time bands

3.2.2 Feature engineering

Density-based features

The density of POIs is one of the first feature types of interest for our model (Cervero and Kockelman, 1997), which represents the presence intensity of amenities or transport nodes of each type in an area that would attract visitors (in our case, coming by public transport). We expect a positive correlation between the density of POIs of all types to the total number of arrivals. An area with a high concentration of retail within a 15-minute walk is expected to be more attractive than a similar-sized area with a low concentration of retail. Similarly, an area with a high concentration of transport facilities, such as stations and bus stops, is expected to accommodate a higher inflow volume of visitors. However, feature importance and spatial nonstationarity (different impacts on the target variable depending on the location) in this regard will be factors of interest as we interpret the model predictions post-training.

All density-related features—i.e., population, amenity POI density and transport node POI density—are formulated by aggregating their values for each spatial unit (isochrone) by type, resulting in a set of **16 features** (See Figure 3.8)

Diversity-based features

Diversity of Amenity POIs according to Cervero and Kockelman (1997) is also an important factor that may influence the attractiveness of an urban area for certain types of non-commute travel demand. The implication from their conclusion is that between two areas, with all else being equal, the area that is more diverse in the types of amenities it provides will be more attractive on average because of their vibrancy and the possibility it can afford visitors to accomplish different tasks in one trip.

What was left under-explored by literature is the diversity of public transport facilities present in a spatial unit of analysis, which is an important factor that may influence the attractiveness of a destination area for non-commute travel demand. Another reason is the possibility of capturing the modal interchange behaviour of passengers whereby they alight in a certain area only to switch between different means in order to reach their final destination elsewhere. Between two areas with the same amenity POI profile, the area that can be reached with multiple bus routes and several rail services may be more attractive to visitors either because it is well-connected or because it

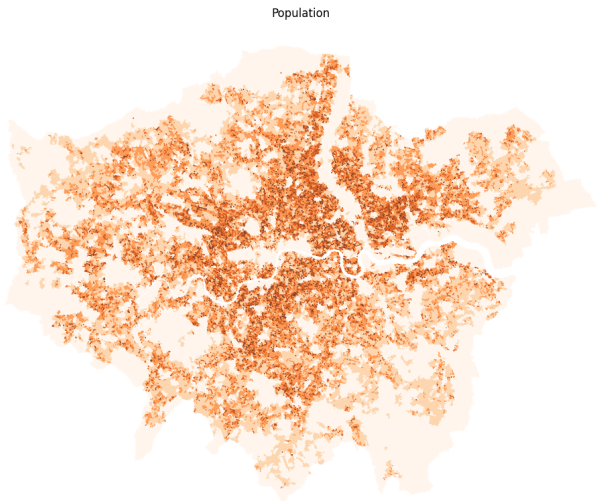

Figure 3.7: Population density (UK Census 2011)

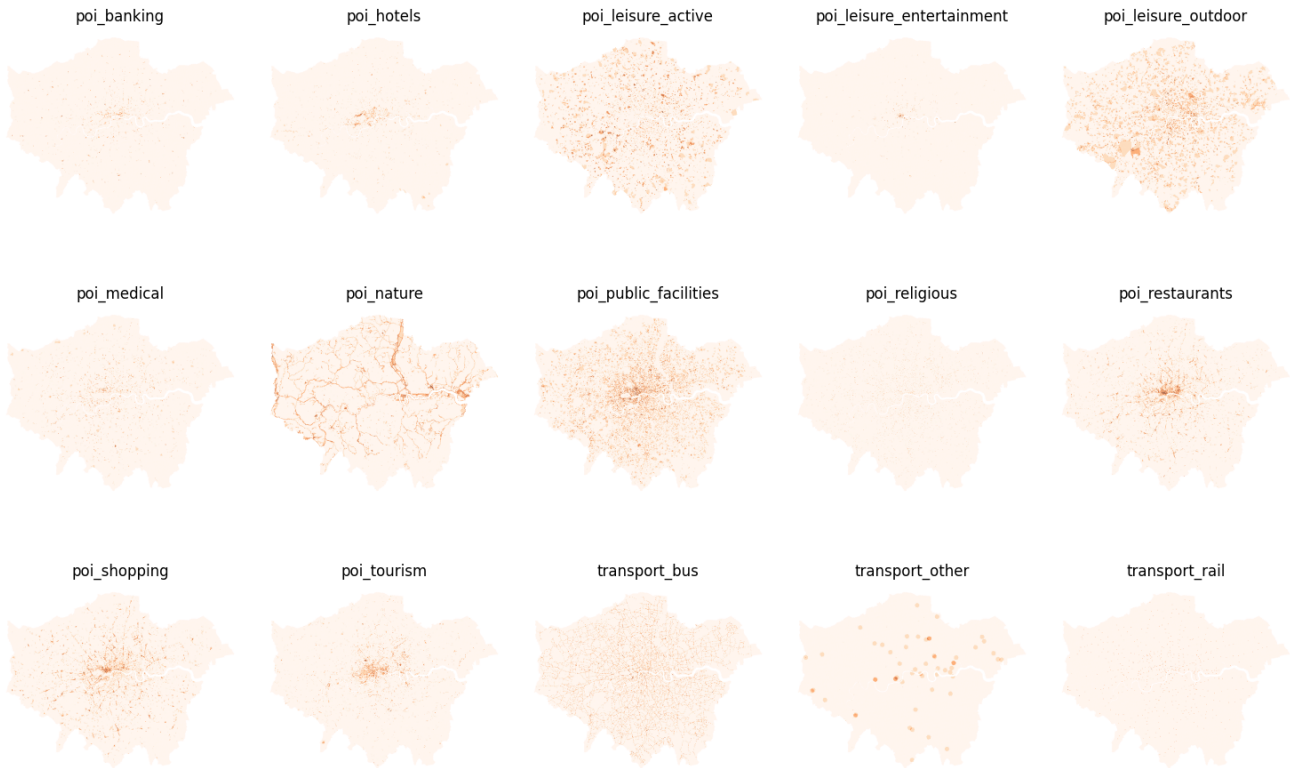

Figure 3.8: Amenity and Transport POI density by type (Source: OSM)

provides opportunities for passengers to switch between different modes of transport. Since there is no origin-destination matrix available for the study area, its inclusion helps the model training disentangle the impact of modal interchange behaviour on the target variable versus the impact of the presence of amenities.

Cervero and Kockelman (1997) conceptualised diversity in the form of *entropy*. A concept originated in information theory, entropy is the average amount of information conveyed by an event when considering all possible outcomes. Batty et al. (2014) further adapted it to convey complexity in a spatial system. Interestingly, they also considered the increasing spatial entropy to mean a trade-off between the density of amenity types and the unique types of amenity present, which stands in competition with other density-based features created. Therefore, the inclusion of entropy as a measure of POI diversity may improve the model prediction accuracy. We intend to replicate the approach used by both of the cited works above and adopt the formulation of entropy introduced by Shannon (1948) to calculate the diversity index of amenity POIs in each spatial unit. The diversity index is to be calculated as follows:

$$H = -\sum_{i=1}^{n} p_i \ln p_i \tag{3.1}$$

where p_i is the proportion of the i-th POI class in the total number of POIs in the spatial unit, and n is the total number of POI classes being considered. The entropy index ranges from 0 to $\ln n$, where 0 indicates no diversity and $\ln n$ indicates maximum diversity, which is then normalised to the 0-1 range. Based on this definition, we will calculate **3 features** related to the diversity of amenity POIs and transport POIs, further explained as follows:

- *Amenity POI*: 12 amenity POI classes are considered for the calculation of the diversity index for each spatial unit (see Table A.2). Looking at the 10% highest values of the diversity index in Figure 3.9, we can see that the most diverse areas are located throughout the Greater London area, which much of urban studies literature links to activity centres or neighbourhoods. As a whole, the distribution of the diversity index is right-skewed, meaning we observe a generally higher diversity of amenities in Greater London than a normal distribution would suggest.
- *Transport POI diversity by mode*: 3 transport POI classes are considered for the calculation of the diversity index for each spatial unit (see Table A.2). The top 10% values of the diversity index are mapped in Figure 3.9, showing that the most diverse areas are located around important transport intermodal hubs, such as national rail stations, airports, and ferry piers.
- *Transport POI diversity by route option*: A more granular diversity index is calculated for the transport POIs, considering the different bus routes and train services the transport POIs in each spatial unit serve. The top 10% values of the diversity index are mapped in Figure 3.9, showing that the most transport-option diverse areas are not only intermodal hubs but also areas around trunk routes where many bus routes or rail services converge.

The inclusion of both transport diversity indices aims to provide complementary information to the model, as they capture different aspects of the transport network that may contribute to the attractiveness of an area for non-commute travel demand. High transport mode diversity indicates a high possibility that the area facilitates modal interchange for onward travel (perhaps long-distance intercity), whereas high transport route diversity, in addition to that, indicates a high possibility that the area is well-connected with good frequency and thus more easily reachable. Zooming into a specific area, such as the Hammersmith and Fulham borough, we can see that there

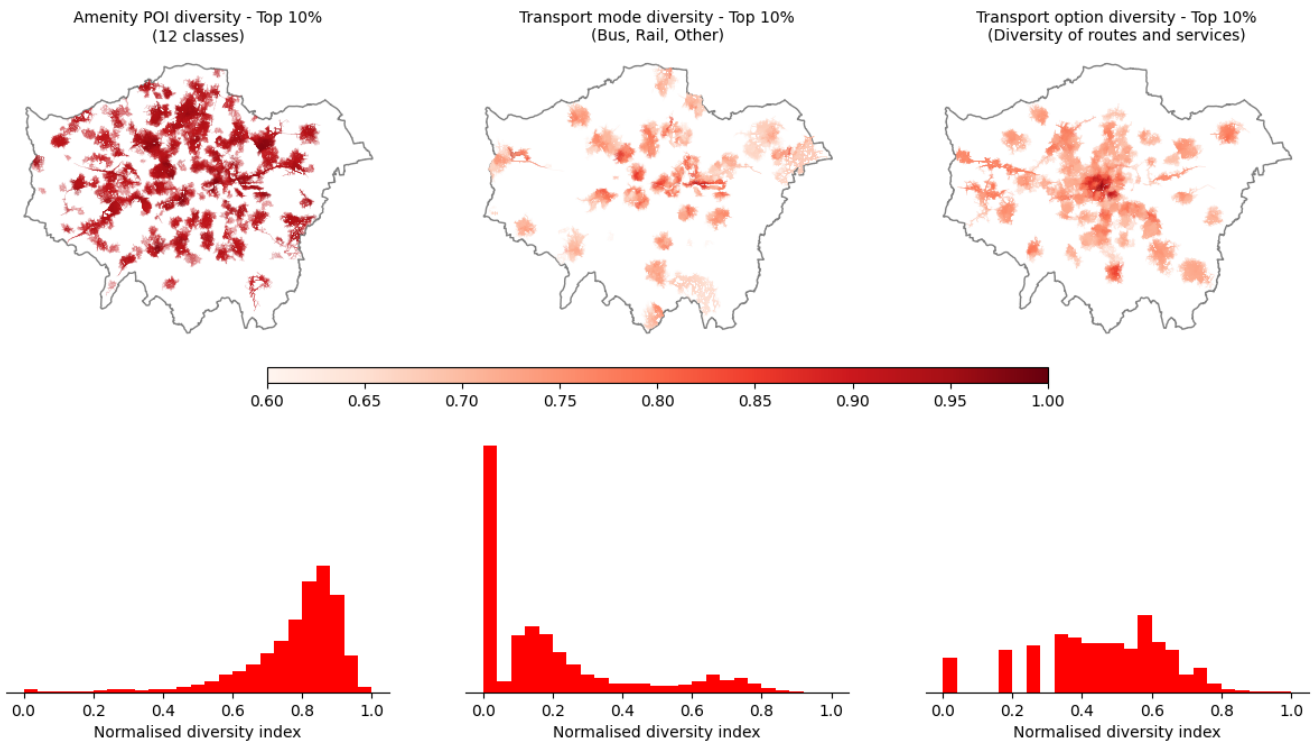

Figure 3.9: Normalised diversity indices for Amenity POIs, Transport POIs by mode and by service. *Top*: Mapping top 10% values, *Bottom*: Distribution of diversity indices

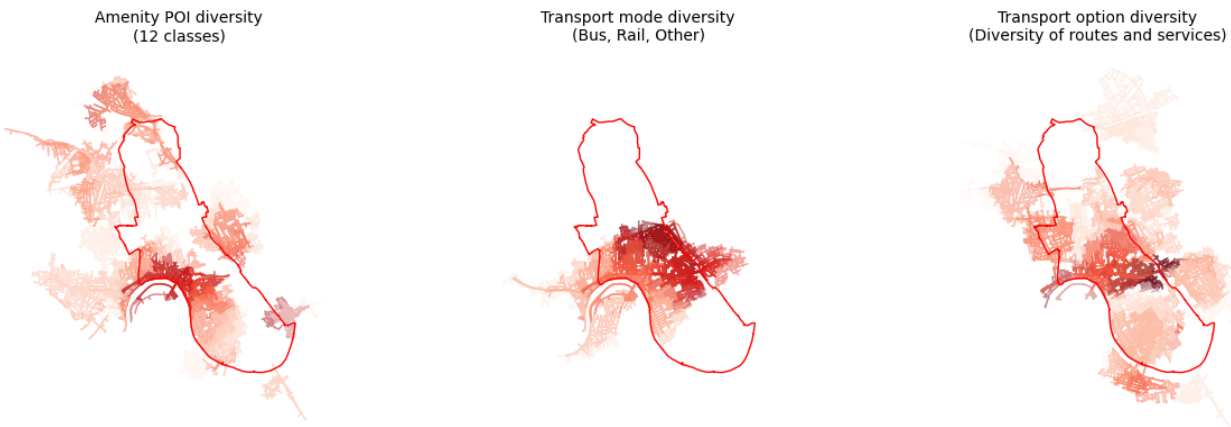

Figure 3.10: Top 10% values of three diversity indices
Borough: Hammersmith and Fulham

is a slight divergence between the two transport diversity indices (Figure 3.10). The Shapherd's Bus area has a higher transport mode diversity index thanks to the existence of national rail services (Southern Railways) beyond Tube and busses, while the area around Earl's Court has a higher transport option diversity index due to the presence of multiple bus routes converging in the area.

Centrality-based features

Whereas the public transport facility diversity index, previously discussed, captures the variety of transport options available in an area, it obscures the accessibility aspect in relation to other

areas in the region of interest using the public transport network. It adds yet another layer of information to the model, as it is expected to capture the importance of a transport node in the network, which may influence the attractiveness of an area for non-commute travel demand. For example, between two areas with similar diversity of transport options, the area with a higher centrality measure may be more attractive to visitors because it is more 'central' in the network.

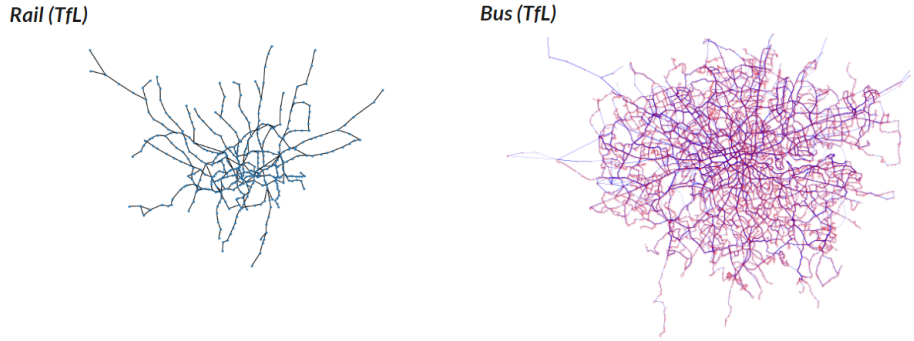

Figure 3.11: Network graphs of TfL Rail (left) and Bus (right) networks (Source: TfL)

To achieve that, first, we construct the network graph of the TfL Bus and the TfL Tube and Rail networks (See Figure 3.11). The nodes in the network represent the transport nodes (stations and stops) in the network, and the edges represent the routes. Due to the lack of data related to travel time or frequency, we abstain from combining the two networks into a multi-layered network, which would have allowed us to consider the whole public transport network as one integrated one. Instead, we will calculate the following centrality measures for all the nodes in the Rail and Bus networks separately, resulting in a set of 6 node attributes:

- *Degree centrality*: The number of direct connections a node has in the network. A high degree of centrality indicates that the node is well-connected to other nodes in the network (i.e., many connections available).
- *Closeness centrality*: The average shortest path length from a node to all other nodes in the network. A high closeness centrality indicates that the node is close to other nodes in the network (i.e., centrally located).
- *Betweenness centrality*: The number of shortest paths that pass through a node in the network. A high betweenness centrality indicates that the node is an important bridge between other nodes in the network (i.e., transport hubs).

Finally, in order to incorporate calculated attributes of the nodes into **6 features** of centrality for the spatial units to train our model, we will consider the nodes that are in each of the spatial units and extract the maximum attributes among them. In other words, the centrality feature of a spatial unit has the value of the rail or bus transport node with the highest centrality measure. The mapping of the centrality measures to the spatial units is illustrated in Figure 3.12.

3.2.3 Incorporating spatial lags as a feature

In summary, the feature engineering process results in a total of **25 features** that will be used to train the model to predict the (log-transformed) total daily arrivals at a transport node or bus stop in a given spatial unit. The features are summarised in Table 3.1.

When working with machine learning models with data that exist in a closed spatial system such as the Greater London area, it is important to consider the spatial autocorrelation in the

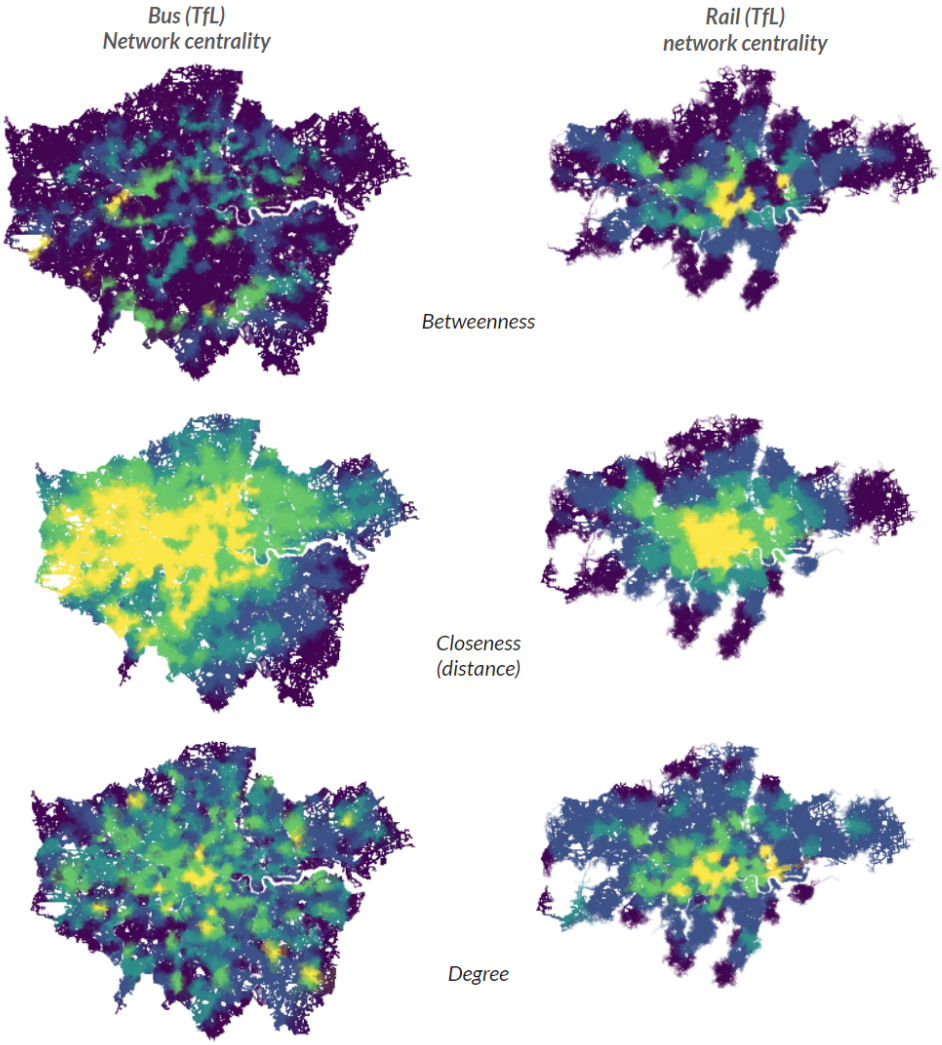

Figure 3.12: Mapping centrality measures of spatial units (isochrones) on Rail (left) and Bus (right) networks

No.	Feature
1	Population density
2-13	Amenity POI density by type (12)
14-16	Transport POI density by type (3)
17	Amenity POI diversity index
18	Transport POI diversity index by mode
19	Transport POI diversity index by route option
20-25	Bus and Rail centrality measures (Degree, Closeness, Betweenness) (6)

Table 3.1: Summary of the 25 main features engineered for the model

target variables. Spatial autocorrelation refers to the phenomenon where the value of a variable at one location is correlated with the value of the same variable at nearby locations. In the context of our analysis, this means that the number of passengers arriving at a transport node or alighting from a bus at one isochrone is likely to be correlated with the number of passengers arriving at

nearby isochrones. Although deep learning models' parameters and hyperparameters can be made more complex and tuned to reach a high degree of prediction accuracy without considering the spatial aspect of the dataset and can practically be considered non-parametric (i.e., not assuming the variables to have a specific distribution), accounting for potential spatial autocorrelation could reduce the probability of biased outcomes even with a low-complexity and high-interpretability architecture, e.g. tree-based (Meyer et al., 2019).

The existence of spatial autocorrelation in the target variables is confirmed by the Local Indicators of Spatial Association (LISA) analysis, which shows the presence of spatial clusters in the target variables (Figure 3.13), using Local Moran's I indicators. High-High and Low-Low clusters of the total daily arrivals with public transport are respectively present in the inner core and the outer edge. Unless spatial relational information among these spatial units is made apparent to the model during tuning and evaluation, we are bound to see low accuracy, especially in these localities that make up more than 50% of the study area.

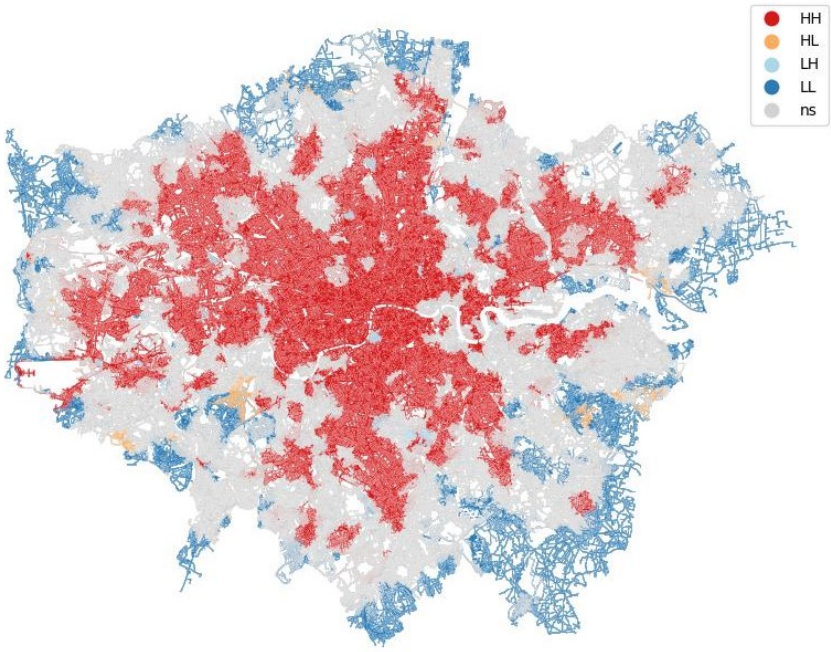

Figure 3.13: LISA cluster map of Total arrivals showing local sptial autocorrelation
Global Moran's I: 0.742 (High spatial autocorrelation)

To address this issue, Liu, Kounadi, and Zurita-Milla (2022) has proposed incorporating spatial lag into the model as a feature. More specifically, the authors have concluded that when spatial features were induced in the random forest models, the global spatial autocorrelation among the predictions' residuals was significantly reduced, also leading to higher validation accuracy. For our analysis, the spatial lag feature of each spatial unit is calculated as the average number of passengers arriving at neighbouring isochrones[1]. The creation of a **Spatial Lag features** is replicated for the total daily arrival counts, as well as for each time band, resulting in a set of additional features to be included to train the model in predicting the corresponding target variable.

[1]The choice of the distance band of 750m was calibrated with the objective to ensure all spatial units have neighbours, i.e., there are no islands.

No.	Target variable	Main Features		Spatial Lag Feature
1	Total arrivals	Population, amenity and accessibility profile (25 features)	+	Spatial Lag - Total arrivals
2	Morning arrivals			Spatial Lag - Morning arrivals
3	Midday arrivals			Spatial Lag - Midday arrivals
4	Evening arrivals			Spatial Lag - Evening arrivals
5	Late arrivals			Spatial Lag - Late arrivals

Table 3.2: Mapping spatial lag features to predict corresponding target variables

3.3 Model selection and interpretation

As previously explored through relevant literature, we are faced with an important decision when attempting to model the non-commute travel demand in the Greater London area based on variables related to the built environments, such as the amenity profile and the connectivity profile. More specifically, the choice is between using a machine learning model, which prioritises prediction accuracy and can handle heterogeneous data, and a spatial-statistical one, which prioritises interpretability and requires more stringent requirements of the explanatory variables. Although spatial autocorrelation and spatial heterogeneity are the two well-known effects in spatial analytics and are factored in statistical models such as the Spatial Lag Model or Geographically Weighted Regression, Li (2022) has shown that machine learning models can capture spatial effects if spatial information is provided. The compromise between the two is the use of machine learning explanation techniques, which can provide both high prediction accuracy and interpretability while being spatially aware. Figure 3.2 provides an overview of the model training and interpretation process.

Model selection

Firstly, we will compare the prediction performance of three machine learning models: *Linear Regression, Gradient Boosting, and Neural Network* in predicting Total arrivals. As we are working within a spatial system with a high degree of spatial autocorrelation demonstrated earlier, special attention will be paid to cross-validation in hyperparameter tuning and final evaluation where random cross-validation techniques and random train-test splits are not able to ensure the absence of leakage of information between the training and testing sets which inflates the accuracy score and to avoid undetected overfitting.

More specifically, the model tuning and evaluation process will be carried out using a nested 10-fold cross-validation approach introduced by Schratz et al. (2018), with R^2 as the goodness-of-fit score. The nested cross-validation approach is illustrated in Figure 3.14, where the inner loop is used to tune the hyperparameters of the model, and the outer loop is used to evaluate the model's performance. The spatial 10-fold cross-validation splits for our datasets (see Figure A.1 in Appendix) ensures that the training and test sets in both the inner and the outer folds are spatially distinct to avoid data leakage. The model and its hyperparameters with the highest R^2 score will be selected as the best model to fit the entire dataset for further interpretation[2].

[2]Although the model tuning and selection process is only performed on Total Arrivals as the target variable for simplicity, the chosen model will be applied to predict arrivals by time bands.

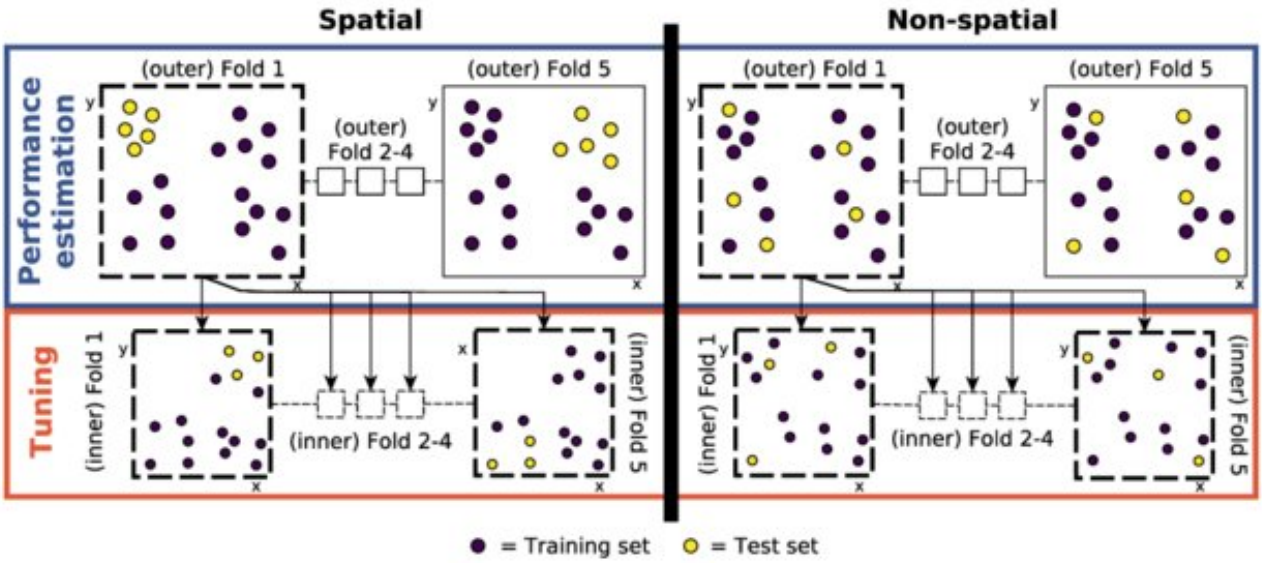

Figure 3.14: Spatial vs. non-spatial nested cross-validation using 5-folds for hyperparameter tuning and performance evaluation. This technique helps ensure the training and test sets in both the inner and the outer folds are spatially clustered to avoid data leakage.
Source: Schratz et al. (2018)

Model interpretation

SHAP, or SHapley Additive exPlanations, is a unified approach to explain the output of any machine learning model. SHAP connects game theory with local explanations (Lundberg and Lee, 2017). Rooted in the work on game theory of Lloyd Shapley, SHAP tackles the challenge of fair credit allocation within a cooperative game. In this reimagined context, features in our data become the players, the set of features used for training represents the coalition, and a single model prediction is the game. This SHAP value quantifies the average marginal contribution of a feature relative to a specified baseline, providing a nuanced understanding of each feature's impact on the model's predictions at the local level, i.e., for each individual prediction.

Among other properties contributing to its popularity in Explainable AI research, it is model-agnostic and additive, meaning that the sum of the SHAP values for all features is equal to the difference between the model's prediction and the baseline prediction. While computation was initially a barrier, approaches like Kernel SHAP and Tree SHAP made available in the SHAP package have made estimation more efficient. SHAP's local interpretability and integration with popular machine learning libraries also make it a powerful tool for spatial data where feature attribution can be visualised geographically. These properties of SHAP allow us to derive our methodology to interpret the model's prediction in the context of the features engineered to answer the research questions:

- SHAP-based *global feature importance* will be used to identify the most important features (amenity and connectivity profiles) to the model's prediction of inflow travel demand, allowing us to understand the general patterns and relationships between the features and the target variables. Since it is important that the model in question attains a high level of accuracy before the SHAP values can be considered insightful, due care will be taken to ensure that the model is well-tuned and evaluated.

- SHAP-based *local feature importance* will be used to interpret the model's prediction for individual spatial units, allowing further investigation into identifying activity centres, transport hubs, and other spatial patterns in the data that can be visualised. More specifically, we will perform a spatial clustering analysis such as HDBSCAN on the local SHAP values to identify spatial clusters where our main features have high (or low) importance in the model's prediction, effectively identifying areas of high (or low) attractiveness for non-commute travel demand for specific destinations (i.e., hotspots).

4. Results

RQ1: Predicting trip attraction with spatial features

Prediction model evaluation and selection

As previously explained, the choice of a suitable machine learning model is critical as Machine Learning explanation techniques such as SHAP would be less insightful if the predictions produced high inaccuracies. The evaluation results in Table 4.1 show that the Gradient Boosting tree-based model (XGBoost) outperforms the other models in predicting the target variable Total arrivals (log), achieving an R^2 of 0.792, and outperforms itself when predicting the pre-transformed target variable Total arrivals with an R^2 of 0.633. Meanwhile, the simple Linear Regression model performs more poorly in comparison through underfitting (i.e., consistently low accuracy with training data folds), and the Multilayer Perceptron Artificial Neural Network model also performs poorly but due to overfitting (i.e., high accuracy achieved with training data folds but low accuracy with validation data folds). This shows that while XGBoost has long been popular thanks to its fast convergence and high validation accuracy, transforming the highly right-skewed target variable is another way to improve prediction performance.

Model	R^2 **(raw)**	R^2 **(log)**
Linear Regression	0.576	0.697
XGBoost	0.633	**0.792**
ANN	0.526	0.410

Table 4.1: Comparing model R^2 results of 3 models predicting Total arrivals (raw) and Total arrivals (log) using spatial k-fold cross-validation

On managing overfitting risks, XGBoost is a standout among other models also for its built-in L1 and L2 regularisation hyperparameters, which penalise the model for complexity and suppress highly correlated input features. This is particularly important when considering the complex patterns of correlations among our input features, as shown in Figure 4.1. Here we observe a few cohorts of highly correlated features such as Population-Public Facilities-Religious POIs, Entertainment-Restautants-Shopping POIs, or among the different public transport centrality measures. Therefore, with XGBoost, we had the opportunity to further fine-tune two hyperparameters *alpha* and *lambda*—which represent L1 and L2 regularisation thresholds, respectively—to achieve even better results. Similar to previously, nested spatial k-fold cross-validation was used to fine-tune the model and validate the results. The fine-tuned XGBoost model improves R^2 up to **0.823** (an approx. 3pp increase) in predicting total arrivals from the pre-tuned model and

Timeband	R^2	MSE	MAE
Total	**0.823**	0.687	0.570
Morning	**0.809**	0.614	0.571
Midday	**0.828**	0.629	0.549
Evening	**0.819**	0.687	0.593
Late	**0.824**	0.635	0.601

Table 4.2: Finetuned XGBoost model evaluation results for
All target variables (log) using nested spatial k-fold cross-validation

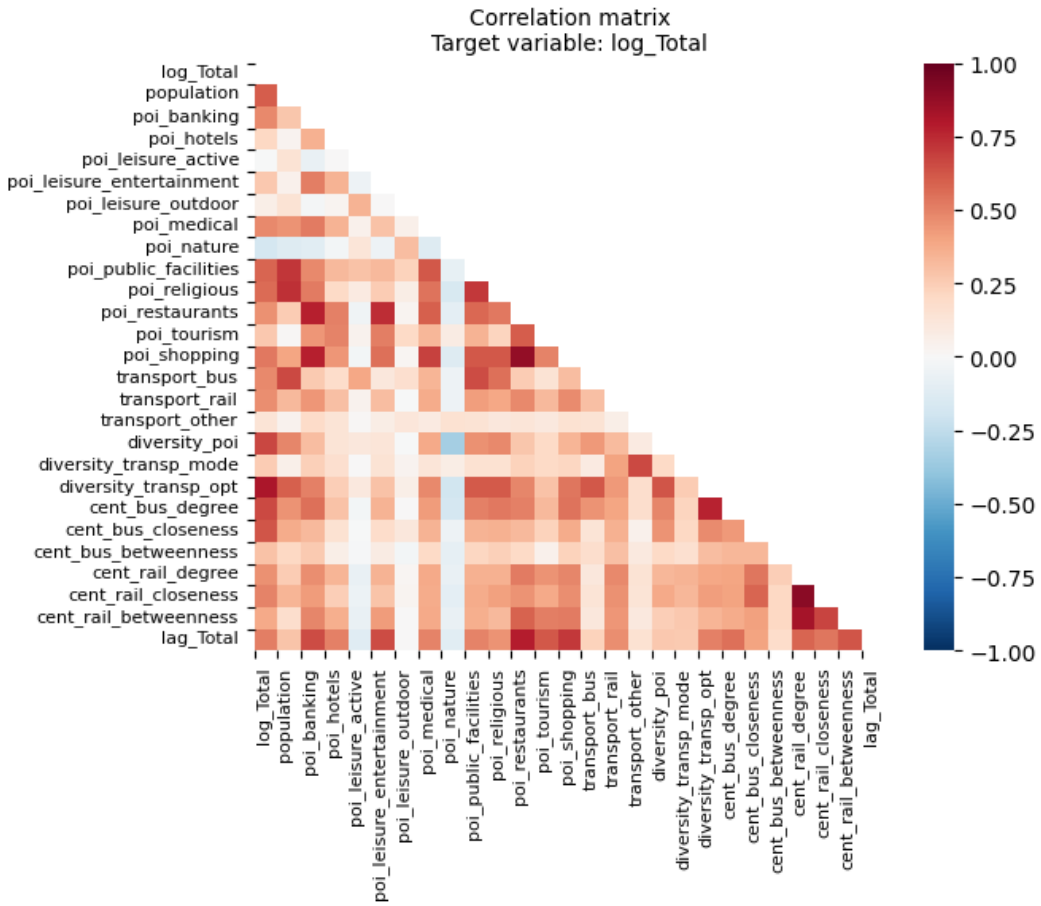

Figure 4.1: Correlation matrix of the input features and target variable Total arrivals (log)

performs consistently well in predicting arrivals by time band, as shown in Table 4.2.

The results suggest that the fine-tuned XGBoost model is fit for purpose and enables us to address the first research question: ztextitBuilt environment features are adequate predictors of public transport inflow demand in London to a high degree of accuracy. This has implications for urban analytics literature and is important for urban planners and policymakers, as it provides a tool that can be used to predict public transport demand and plan for future infrastructure development that relies entirely on accessible data sources such as OSM and the UK Census.

Before further analysis, it is important to recognise where the model falls short and what that may tell us about its potential lack of generalisability. Figure 4.2 shows the distribution

of observations whose prediction residuals are 'outliers', i.e., the target variables are significantly overpredicted or underpredicted. These are mostly concentrated in Outer London where poor predictions go in both directions. While it is not easy to unravel, possible root causes include:

- *Low Spatial Autocorrelation:* There is some consistency between areas with low spatial autocorrelation of the target variable (see Figure 3.13, grey regions), and where the model predicts more poorly. This may indicate incorporating spatial lag as a feature does improve the model's performance only where the spatial autocorrelation is present. In its absence, the model has a higher chance of poorer predictions based on other spatial features alone.
- *Omitted features:* It is possible that adding variables such as car ownership and socioeconomic factors can help the model distinguish two hypothetical areas with the same built environment profiles but different public transport usage. This is especially true in Outer London, where car ownership is higher, and public transport services are less frequent.
- *Incomplete target variable:* It is probable that a spatial unit with its built environment profile does have the predicted public transport trip attractiveness in reality. However, since only data from TfL-operated services are considered, omitting trip data from other operators more prevalent in Outer London and especially south of the Thames, it may show up as having an apparent overprediction.

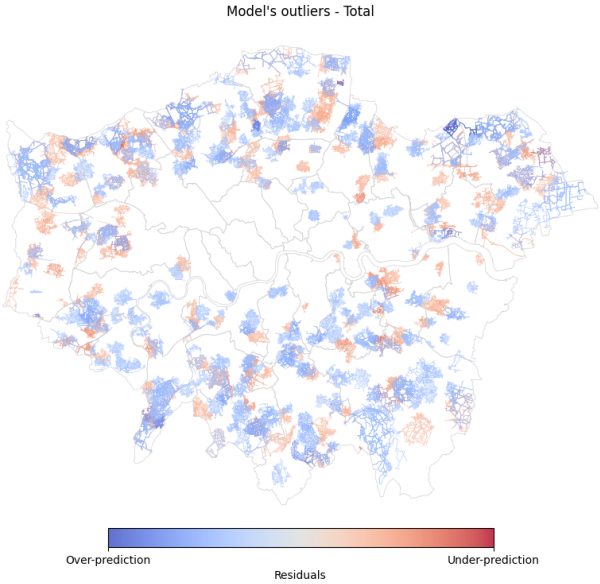

Figure 4.2: Observations with significant overprediction and underprediction (N=576 of 16,890) Fine-tuned XGBoost model, Total arrivals (log)

Nonetheless, these 'outliers' account for less than 3.5% of the total observations, and the model's performance is still satisfactory in explaining the large majority of the spatial units in the Greater London area. Of course, due care is advised when applying the model directly to other localities, and further developments with more comprehensive datasets and additional features may be necessary.

Global feature importance

After fitting the fine-tuned XGBoost model to the entire dataset, we applied interventional SHAP to obtain the SHAP values for each feature and observation. This allows us to evaluate the

importance of each feature in predicting the target variable and understand the direction of the relationship between the feature and the target variable.

Figure 4.3 shows the summary of feature importance for the target variable Total arrivals (log) in order. First, it is evident that the connectivity profile features have outsized importance with the *transport route option diversity index (diversity_transp_opt)* on top in predicting trip attractiveness. This is consistent with the reality that people tend to alight/exit at transport-rich areas where they have multiple bus routes or rail services for onward journeys, representing the interchange behaviour. The second most important feature is the spatial lag of the target variable (i.e., mean of target variables of neighbouring), which partially corroborates its influence to the model's performance hypothesised earlier when analysing. The most important amenity POI feature is *shopping POIs*, which is also consistently important across all time bands (see Figure A.2 in Appendix), followed by *banking POIs*, *active leisure POIs*, *medical POIs*, *restaurant POIs*, *nature POIs*, and *outdoor leisure POIs*. This suggests that for the whole of Greater London, total arrivals as a proxy for trip attractiveness are driven mainly by, first and foremost, the need to interchange for onward travel and, secondly, by the availability of certain types of amenities.

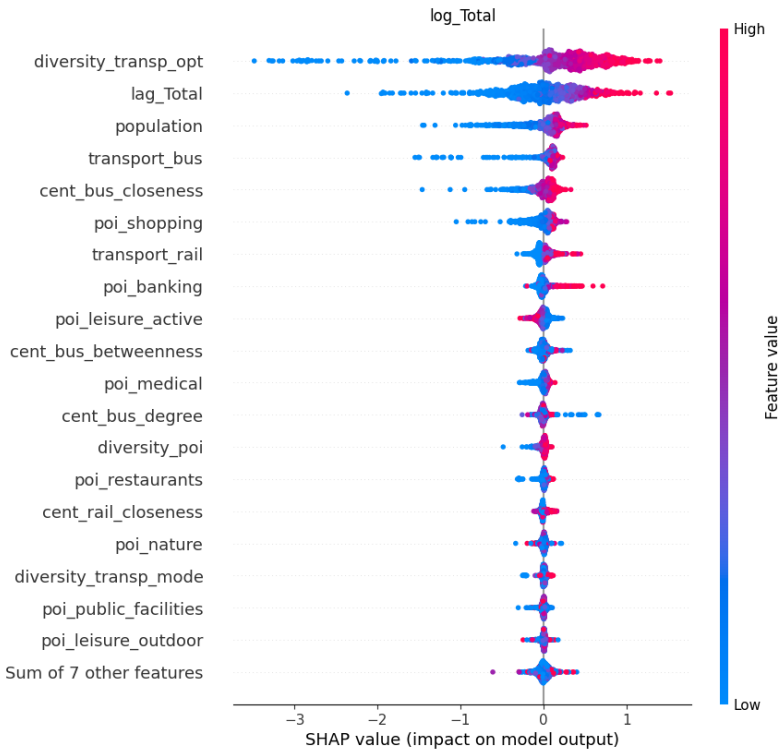

Figure 4.3: Summary of feature importance based on SHAP values of a sample. Red-Blue scale represents the feature value, and the x-axis represents the SHAP value, i.e., impact on final model prediction of Total arrivals (log)

This may seem counterintuitive at first that banking POIs are deemed more important in predicting trip attractiveness on a Saturday, but it is important to consider (1) the correlation between the input features as previously observed and (2) how the XGBoost model learns. Although regularisation is in place to suppress highly correlated features to cause overfitting, it does not drop them from the model. Consequently, when SHAP is applied for a retroactive explanation, highly correlated features will show comparable importance. In such cases, it is, therefore, more insightful to consider the SHAP values of the correlated features in clusters, as shown in Figure 4.4. We can see that there are differences from the previous graph that shows the order of

importance of the features (or feature clusters) in explaining trip attraction to an area as follows:

1. Transport route option diversity index
2. Spatial Lag of Total arrivals (log)
3. Population + Public Facilities
4. Bus stop density + Bus stop closeness centrality (interchange hubs)
5. Density of Shopping, Banking and Restaurants POIs
6. Rail station and other transport node density
7. Others

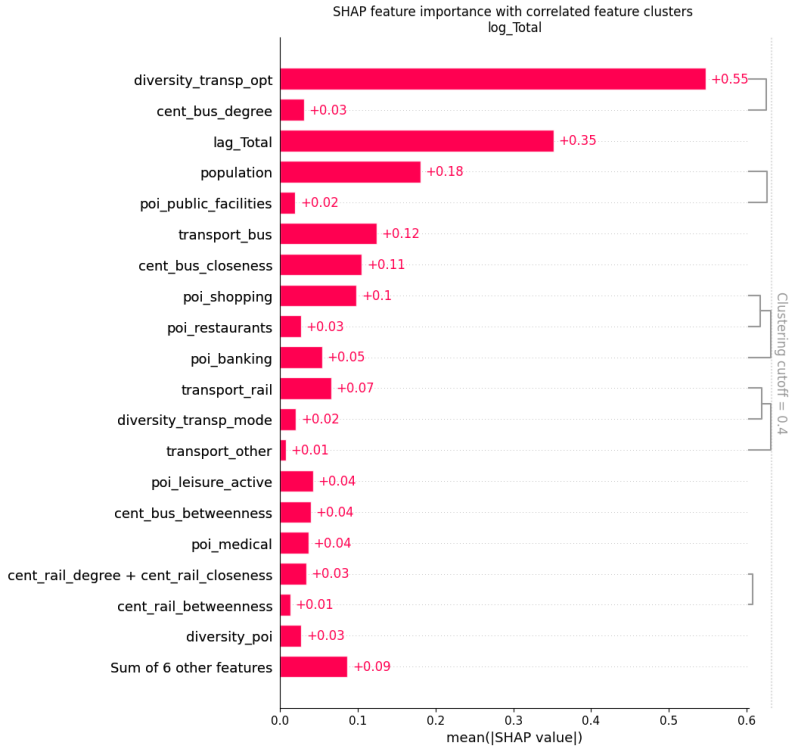

Figure 4.4: Feature importance in predicting Total arrivals (log)
Features clustered by correlation and ordered by mean absolute SHAP values

In summary, as it pertains to our first research question, the global feature importance analysis shows that built environment features are indeed important predictors of public transport inflow demand in London and that the XGBoost model is able to capture the complex relationships between these features and the target variable with high accuracy (¿80%). There are also opportunities to improve the prediction performance by addressing possibly omitted spatial and non-spatial features. The high importance of the spatial lag feature is both an advantage and a shortcoming: While it drastically improves the model's performance overall, it penalises the model in areas where the spatial autocorrelation is low, reducing its generalisability to predicting the trip attractiveness of an area in isolation without considering its neighbours. Finally, highly correlated features should be considered in clusters when analysing importance to avoid interpreting undue importance to a single feature.

RQ2: Identifying activity hotspots

One of the highlights of SHAP is its ability to provide local explanations for individual predictions. In the case of a spatial system, these local explanations provide additional insights on spatial phenomena that are not well captured in a global feature importance analysis, namely spatial heterogeneity (i.e., varying impact of the target variable across different spatial units) and spatial non-stationarity (i.e., varying impact of an input feature on the target variable across different spatial units). SHAP's additive property allows us also to decompose the prediction of a spatial unit into the sum of the SHAP values of each feature and exact the contribution of each feature to the prediction.

Figure 4.5 shows the local SHAP values for two spatial units, Kew Gardens and Covent Garden, for a closer look. They represent two distinct localities with supposedly different amenity, transportation and trip attractiveness profiles. Kew Gardens is a relatively quiet but well-connected residential area in the public transport network, also known for its many renowned parks and gardens, while Covent Garden is a bustling commercial and entertainment district with high connectivity in the public transport network. The local SHAP values for Kew Gardens show a higher contribution of the population density and POI classes that are associated with neighbourhood amenities (stores, clinics, etc.) to the prediction of total arrivals, while the local SHAP values for Covent Garden show a higher contribution of POI classes such as restaurants, entertainment venues, etc. This is consistent with the expected profiles of the two areas and demonstrates the ability of SHAP to provide local explanations for individual predictions. Finally, in both cases, in line with global feature importance, the transport route option index's contribution is among the top, indicating that with the station exit and bus alighting dataset at hand, the interchange behaviour is still a significant predictor regardless of where in London the spatial unit is located.

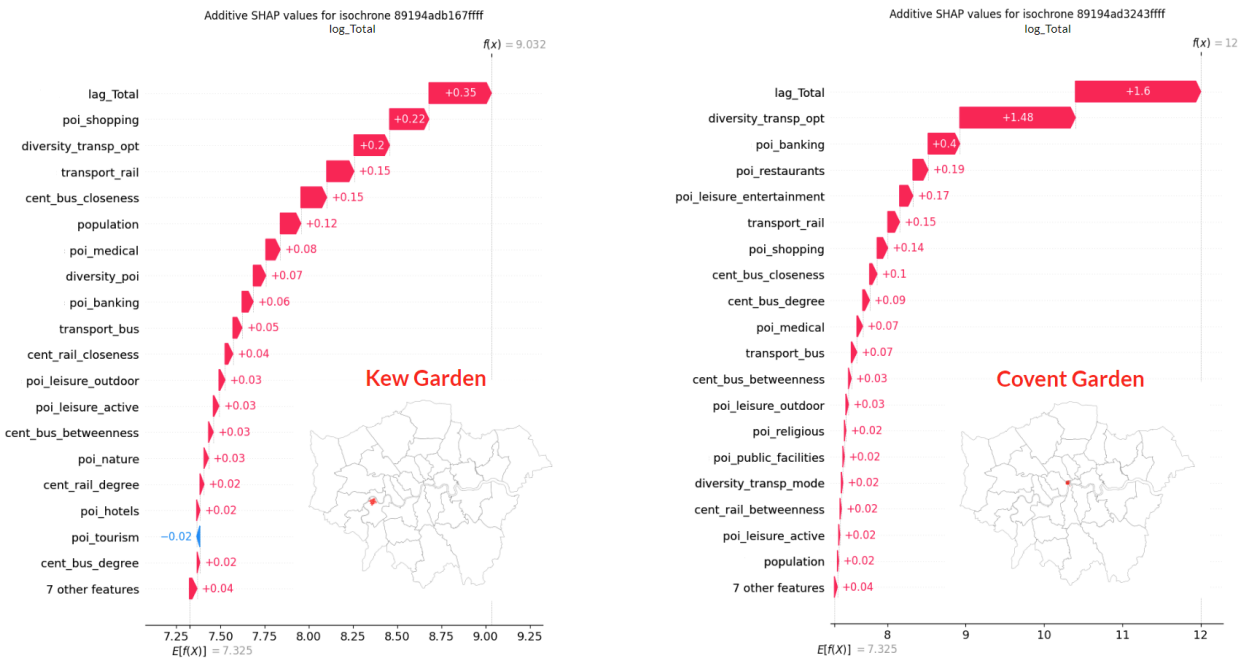

Figure 4.5: Local SHAP values for 2 spatial units
Kew Gardens and Covent Garden

By mapping the local SHAP values for all spatial units in the Greater London area, we can visualise the differing impact of the input features on the target variable across the area. Figure 4.6 shows the local SHAP values for two features, Outdoor leisure POIs and Entertainment leisure

POIs, for the target variable Total arrivals (log). The map shows that the impact of these features on the target variable is not uniform across space, with some areas showing a positive impact (red) and others showing a negative impact (blue). Central London understandably commands the highest contributions of entertainment POIs to the prediction of trip attractiveness, but it also shows secondary areas of high impact in the outer boroughs, such as Brixton and Haringey. Outdoor leisure POIs, on the other hand, show a more dispersed impact citywide with a hotspot around Richmond.

On the temporal dimension across time bands, we can also observe the varying impact of the input features on the target variable across different time bands. Figure 4.7 depicts local SHAP values spatial distribution throughout the day normalised by the target variable (arrival by time band) to enable comparison. Entertainment POIs in Central London have the highest trip attractiveness in the Midday and especially Evening time bands. In comparison, Outdoor leisure POIs show a more consistent impact across time bands with a notable increase in the intensity of certain hotspots and coldspots in the Late time band.

While it is not immediately clear why this surfaces, it reveals a shortcoming in the methodology when examining the temporal effect of the feature contribution to trip attractiveness among different features. According to global feature importance for the Late time band, *outdoor leisure POIs* seem to have an increased impact on the prediction contrary to the expectation that no passengers would be visiting outdoor leisure POIs after midnight. Since SHAP's goal is to ultimately approximate how a trained model makes decisions, it may overestimate the importance of some features in the absence of others that might be able to explain the variances. This also means the choice of the model training methodology and the feature selection process so far may not be a one-size-fits-all for temporal datasets, and additional features might be necessary to avoid overinflating the contribution of certain features.

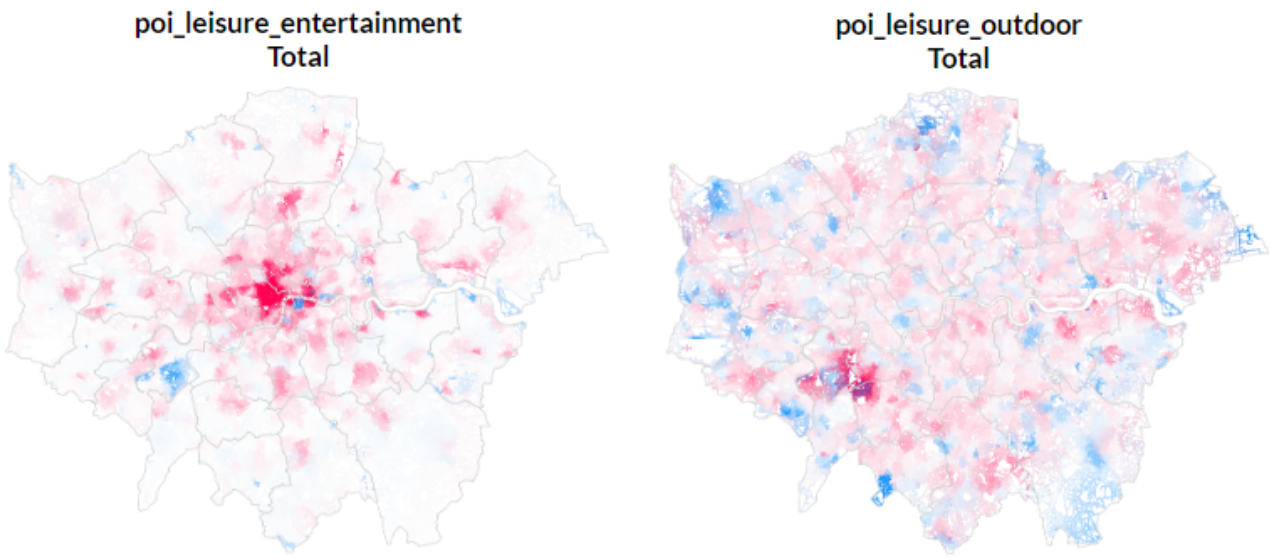

Figure 4.6: Mapping local SHAP values for 2 features: Outdoor leisure POIs and Entertainment leisure POIs. Target: Total arrivals (log)

Finally, with the availability of raw SHAP values also enables other manipulation for specific purposes. For example, with different types of weightage, we can observe the impact of the features on the target variable in a different light. Figure 4.8 shows the comparison of raw and weighted SHAP values for two features, Outdoor leisure POIs and Entertainment leisure POIs, for the target variable Total arrivals (log), weighted in two different ways:

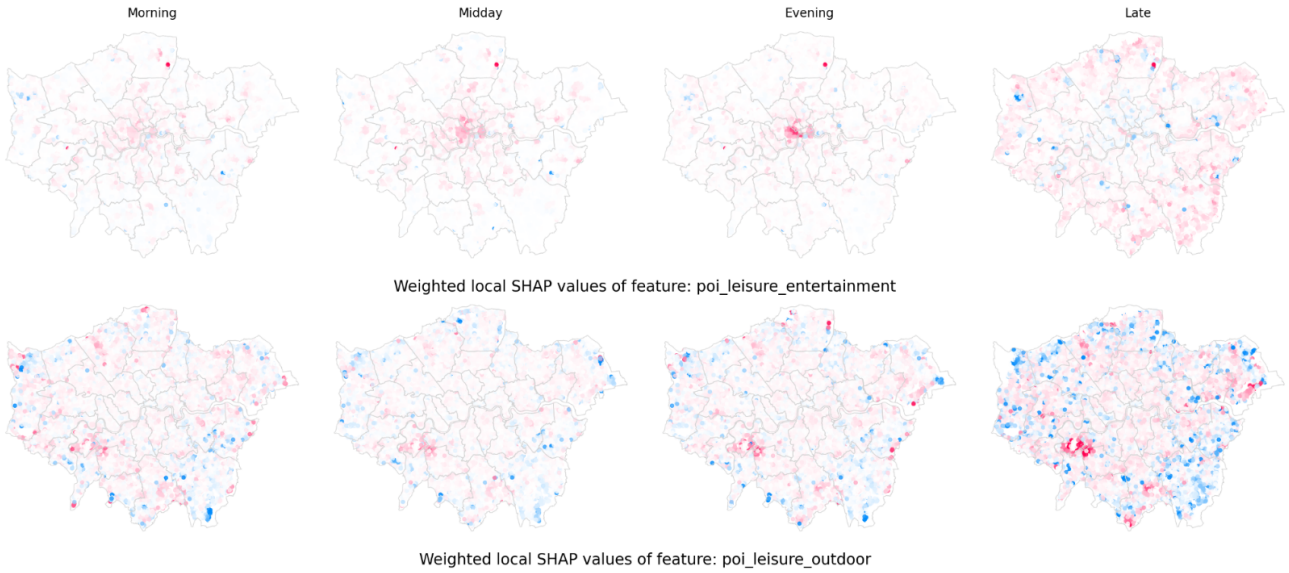

Figure 4.7: Mapping local SHAP values for 2 features: Outdoor leisure POIs and Entertainment leisure POIs, weighted by number of arrivals

1. Weighted by amenity diversity index: This can add more highlights to feature hotspots in areas with many other potential contenders for trip arrivals. Under this treatment, we can see that Entertainment POIs have an even more pronounced impact in Central London and clusters of small areas with high amenity diversity.
2. Weighted by transport route option diversity index: This can add more highlights to feature hotspots in areas that are not well-served by many public transport routes. Under this treatment, we can see that Outdoor POIs on the outskirts of London now have a weighted trip attractiveness, taking into consideration how not as well-connected they are.

Overall, SHAP local interpretability with individual SHAP values for each spatial unit provides a more nuanced understanding of the impact of the input features on the target variable across space and, in our case, time. This is particularly useful for identifying activity hotspots in the Greater London area, as it allows us to pinpoint areas where certain features have a high impact on the prediction of trip attractiveness. Limitations exist, however, in the methodology as the comparative power of SHAP is diminished when many models predicting many target variables are involved, as they are much more adept at enabling comparison of feature importance within the same model to predict the target variable. In our discussion, we will also consider the implications of these findings on urban analytics literature and transport planning.

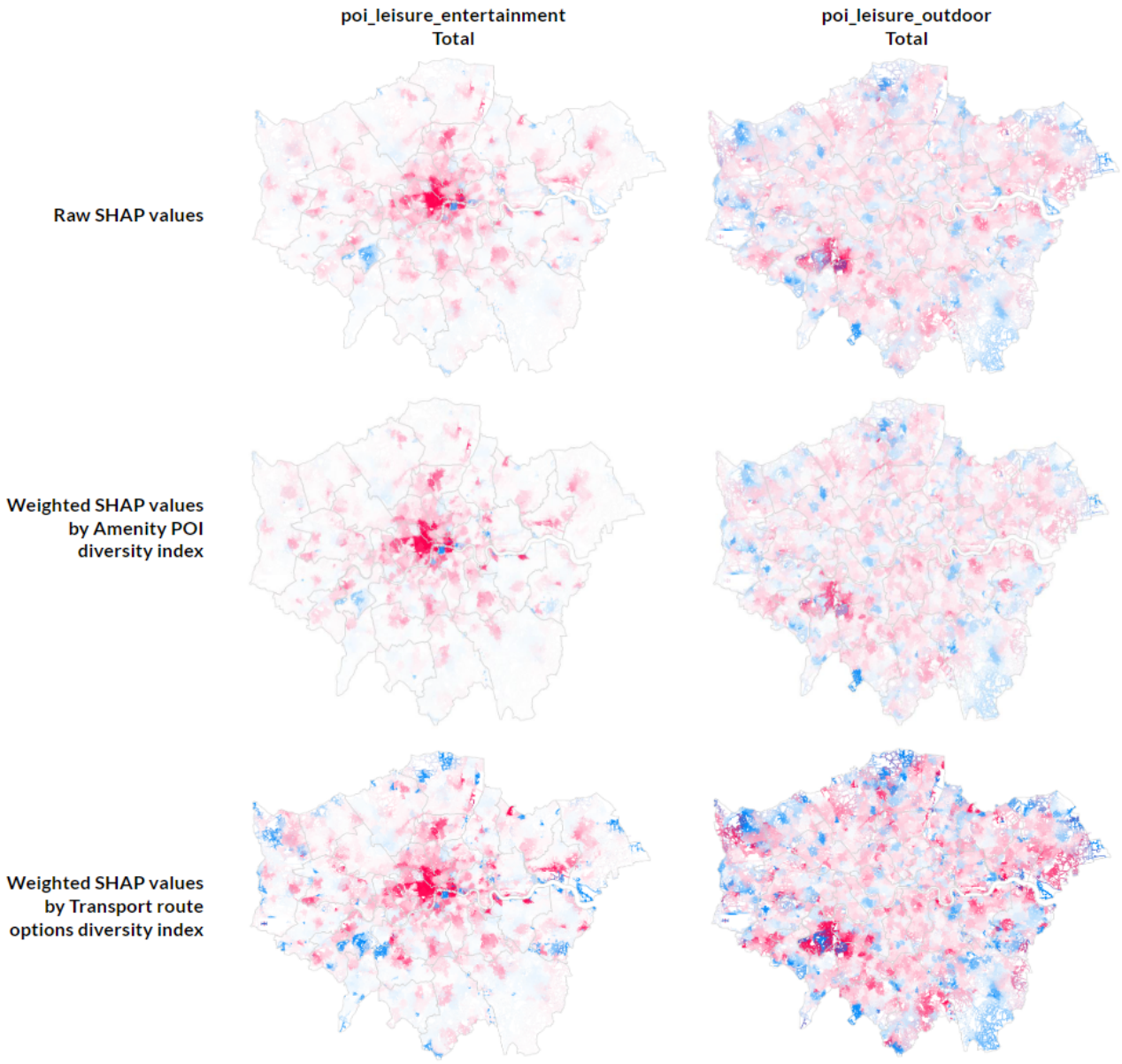

Figure 4.8: Raw vs. weighted SHAP for 2 features
Outdoor leisure POIs and Entertainment leisure POIs
Total arrivals (log)

5. Discussion

To recapitulate our findings thus far: First, we have established that provided with open data on population, amenity points-of-interest density and diversity, and select public transportation accessibility indices, we can predict the total number of arrivals by public transport to a high degree of accuracy using an XGBoost model that is fine-tuned to overcome overfitting due to highly correlated features. The transformation of the target variable and the inclusion of spatial lag features were essential to achieving a high R^2 greater than 0.8 even after already accounting for spatial autocorrelation using nested k-fold cross-validation to avoid inflated performance metrics. Secondly, using SHAP for machine learning model explanation, we managed to extract more sophisticated feature importance indications than originally available with XGBoost. Furthermore, by examining local SHAP explanations, we can identify hotspots for different POI features overall and across different time bands.

One of the benefits of using SHAP values as imperfect approximants to coefficients in a statistical model is their additive nature. This means that we can decompose the model's predictions into the sum of the SHAP values, with which we can control for unwanted effects. For example, without an explicit origin-destination matrix and dwell time data and using only station exits and bus alighting data, it is not immediately possible to say whether a person is exiting at a destination to attend an activity or to interchange to another mode of transport. Using SHAP to extract local importance with its additive property can help consider the impact of the amenity profile of the destination area on the trip attractiveness of the area while controlling for the connectivity profile of the area.

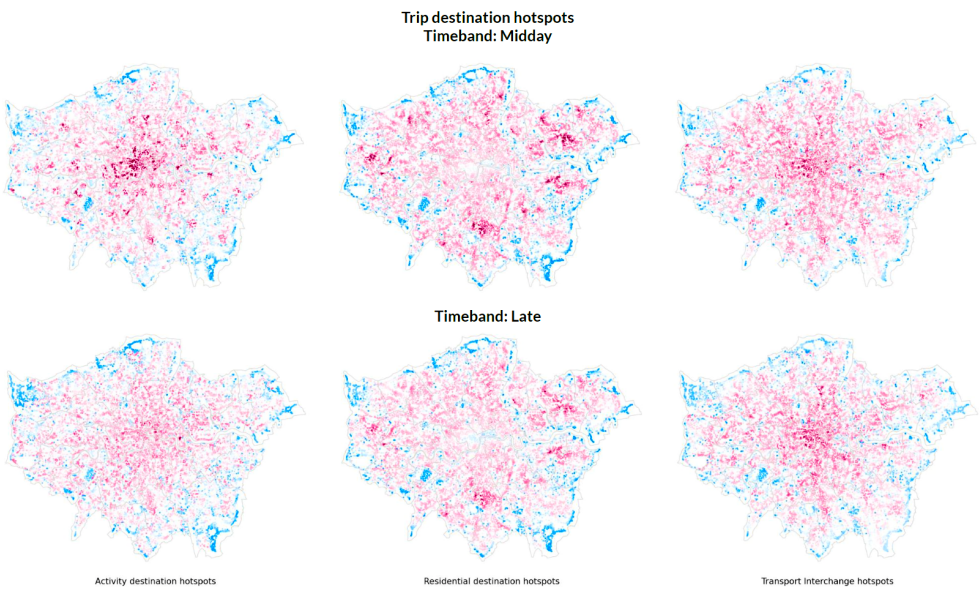

Figure 5.1: Hot spots for destination groups: Amenity vs. Residential vs. Transport Midday vs Late

One way to visualise it is by grouping local SHAP values by feature groups and visualising their spatial distribution. This can be seen in Figure 5.1 with features grouped into (1) amenity-related, (2) population-related, and (3) connectivity-related features. When comparing Midday and Late time bands, we can see that the trip attractiveness of areas because of their connectivity profile is relatively unchanged, whereas the trip attractiveness of areas because of their amenity profile is significantly more pronounced in the Midday time band, despite the fact that connectivity-related features have high global importance. This suggests that we can segment whether an area attracts public arrivals because of their amenities or their connectivity in the network. This is useful for instances where we want to qualify certain popular areas as destination hotspots based on the availability of amenities alone.

Applying spatial clustering such as HDBSCAN of spatial units based on their amenity-related compound SHAP values can effectively surface activity hotspots in a given area in Greater London that are destinations for (non-commute) public transport trips. Figure 5.2 demonstrates the activity hotspot in Wood Green with the highest mean amenity-related SHAP values, made up of 5 constituent spatial units (isochrones). This can be useful for local authorities to identify areas that are popular with public transport users for non-commute activities, such as shopping, dining, or entertainment.

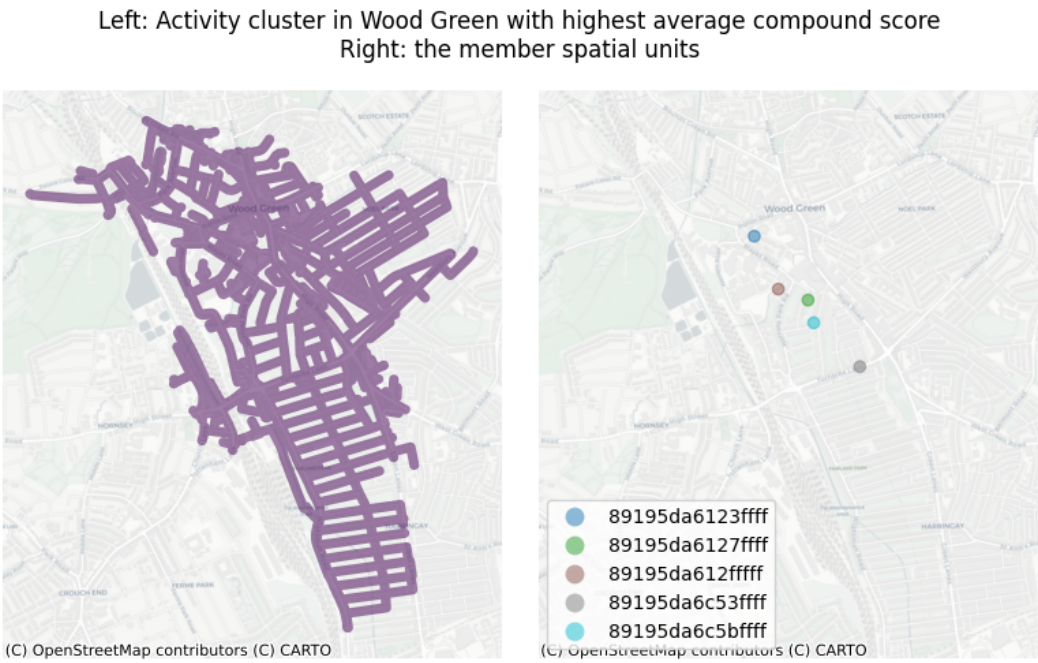

Figure 5.2: Identifying high activity areas in Wood Green

Finally, it is worth acknowledging the limitations of the methodology that affect its generalisability and applicability in certain cases, many of which have been discussed in the previous sections.

- *Missing features*: The analysis of outliers suggests that there are other factors that may be influencing the number of arrivals in an area that are not accounted for in the model. These could include socioeconomic factors, car ownership, or other urban morphological features that affect people's usage of public transport.
- *Generalisability*: The high feature importance of the spatial lag features across the board means that the model may not be able to accurately predict a new observation in isolation

from its neighbours, nor can it be readily generalisable to other cities. Nevertheless, the model can be used to predict the trip attractiveness of a given area in Greater London as the city changes, giving transport planners a useful tool. For example, when a neighbourhood is undergoing redevelopment, we can predict the potential inflow of public transport trips based on the planned addition of amenities or transport facilities.

- *Causality*: However, the lack of consideration for causality means that we cannot infer that the presence of a certain amenity of a certain density will cause more arrivals to a certain area. Rather, we can only observe which amenity or connectivity features among those included in the training of the model are the best predictors of arrivals in an area.
- *POI classification*: The classification of amenity and transport POIs in this analysis adheres to Geofabrik's OSM data classification. Future work could attempt to reclassify these POIs in a deliberate manner to capture the association between certain trip purposes of interest and the types of urban amenities that fulfil them.
- *Types of mobility*: Our analysis uses public transport demand data from one agency (TfL) over a typical Saturday to approximate non-commute public transport trips, meant to remove the effect of employment centres from the analysis as much as possible. However, this does not account for other types of mobility, such as cycling, walking, or driving, which may be more prevalent on weekends, nor does it account for those who work on weekends. Future work could incorporate mode-agnostic mobility data together with trip purpose inferences to exact more precise non-commute trip attractiveness predictions.

6. Conclusion

A. Appendix

A.1 Open data sources

The analysis carried out in this dissertation will utilise the Greater London area as the working case study, serving as an example for further replications of the analysis for other localities. The Office for National Statistics (ONS) defines the Greater London metropolitan area to include the City of London and the 32 London boroughs, with a total area of 1,572 km^2 and a population of 9.4 million people, making it the largest in the United Kingdom. The choice of the Greater London metropolitan area is also motivated by the availability of open data sources from which necessary features could be derived and the complexity of the public transport network that is coupled with rich usage data made available to the public. The data sources used in this analysis are summarised in Table A.1 and detailed below.

	Variables	**Sources**
Public transport arrivals (target)	Tube and Rail station exits Bus alightings	TfL network demand (NUMBAT) TfL network demand (BUSTO)
Population	Population count	UK Census 2021
Amenity Profile	POI count by type POI diversity index	OSM Amenity POIs
Public Transport (PT) Connectivity Profile	PT node count by type PT node diversity index PT network centrality	OSM Transport POIs and TfL

Table A.1: Summary of data used in the analysis and their sources

Transport for London (TfL) Open Data

We use the following GIS datasets from the TfL GIS Open Data Hub and TfL API to derive transport-related features for the study area including density of transport facilities, diversity of transport options, and accessibility in the form of centrality measures in the Rail and Bus networks separately. The networks of interest are defined as follows:

- *Rail network* includes the London Underground, London Overground, the Docklands Light Railway, and the Elizabeth Line, excluding stations and segments that are not managed by TfL or fall outside the Greater London area.

- *Bus network* includes stop locations and route data of all bus and tram services, excluding stops and route segments that are not managed by TfL or within the Greater London area.

TfL also provides annually-updated network demand datasets, which will be used to derive the target variable of our analysis. The datasets are special interest to our analysis are namely *Rail demand data (NUMBAT)* and the *Bus demand data (BUSTO)*, from which we will extract estimated counts of station exits and bus alightings at the station or stop level, respectively. The dataset's temporal granularity is also an important aspect to consider. To address non-commute trips for the analysis, we will specifically use **Saturday average demand**, representing the demand in a typical period without any major citywide disruptions or events[1]. To examine possible variations between different times of day, besides the total daily counts of exits and bus alightings, we will aggregate the data into four larger time bands[2]:

- *Morning* (05:00 - 10:00): Limited mobility expected due to early hours
- *Midday* (10:00 - 19:00): High daytime mobility expected
- *Evening* (19:00 - 00:00): High evening mobility expected
- *Late* (00:00 - 05:00): Minimal mobility expected due to sparser night PT services

It is worth noting that despite the existence of the Oyster Card integrated fare system, TfL does not provide origin-destination data for individuals as open data. Therefore, station exits and bus alightings are not directly associated with the number of passengers traversing through the system but are treated as discrete events. For example, a person who takes a bus to reach a Tube station for onward travel and then exits to arrive at their destination would be counted as two separate data points, one alighting for the bus and one station exit for the Tube. Nevertheless, our model will take this into account when making poredicition and allows us to investigate patterns and the contribution of modal interchange behaviour in our interpretation.

Lastly, we need to acknowledge the particularity of the TfL bus demand dataset (BUSTO). Although passengers do not 'tap off' when alighting from busses, TfL has developed a methodology to estimate the number of passengers alighting at each bus stop based on individuals' next actions and other statistical methods. The methodology is described in detail in the BUSTO User Guide and Data Dictionary

UK Census Data 2021

For the estimated population datasets, we rely on the most recent United Kingdom Census data administered in 2021. The data is ingested at the Output Area (OA) level, which is the smallest geographical unit for which census data is available. Further processing will aggregate the data to the desired spatial unit of analysis.

OpenStreetMap POIs

OpenStreetMap (OSM) is a collaborative mapping project that provides a free and open-source map of the world. The version of OSM data used in this analysis is packaged by an OpenStreetMap

[1]This Saturday travel demand may still include work trips for those working on Saturdays, albeit not concentrated into peak periods as with weekday travel demand. For the scope of this analysis, we will treat these potential work trips as non-commute.

[2]The choice of time band delineation is a compromise between different granularity levels afforded by the TfL rail and the bus demand datasets. There are opportunities to deepen the analysis by further segmenting the time bands into finer intervals if data permits.

community project Geofabrik and provides a snapshot of the OSM database as of June 2024. Geofabrik also standardises the categorisation of POIs' amenity attributes in the OSM database into larger classes[3]. Our analysis will use this default POI classification, extracting 12 amenity POI types. Lastly, apart from amenity POIs, we also make use of the OSM transport POIs point data to derive non-TfL transport-related features, such as national rail stations, ferry piers, long-distance bus stations and airports.

Combined with transport node data acquired from TfL open data sources, all extracted Amenity and Transport POI types of interest are summarised in Table A.2.

	POI Type	**Examples**
Amenity	*Public Facilities*	Post offices, schools, universities, libraries,...
	Medical	Hospitals, clinics, pharmacies,...
	Entertainment	Theatres, Night clubs, cinemas,...
	Outdoors	Parks, playgrounds,...
	Active	Sports centres, swimming pools, stadiums,...
	Restaurants	Restaurants, cafes, pubs, bars,...
	Hotels	Hotels, motels, hostels,...
	Shopping	Supermarkets, bakeries, florists, bookshops, malls,...
	Banking	Banks, ATMs,...
	Tourism	Tourist attractions, museums, monuments, viewpoints,...
	Religious	Churches, mosques, temples,...
	Nature	Riverbanks, beaches, hills,...
Transport	*Bus stops*	Bus stops and stations (TfL)
	Rail stations	TfL rail stations (TfL) and Non-TfL rail stations (OSM)
	Other transport hubs	Airports, ferry piers, long-distance bus stations (OSM)

Table A.2: Summary of Amenity and Transport POI types used in the analysis

A.2 True to Data and True to Model in SHAP technique

One ongoing debate regarding the use of SHAP for explainable is captured by Chen et al. (2020)'s work exploring 'True to Model' versus 'True to Data' explanations. The 'True to Data' approach produces *observational SHAP values* which reflect how a feature contributes to the model's prediction when considering the dependencies and correlations present in the training data, whereas the 'True to Model' approach produces *interventional SHAP values* which reveal how a feature influences the model's prediction if we manipulate it independently, regardless of how it might interact with other features in reality. In our case, as we are more interested in observing feature contribution to the model prediction in a controlled environment rather than observing the relationships between features, we will use the 'True to Model' approach to interpret the model's prediction made possible by the *shap* package.

[3]More information about Geofabrik's OSM data product can be found in Geofabrik OSM data dictionary

A.3 Supplemental Figures

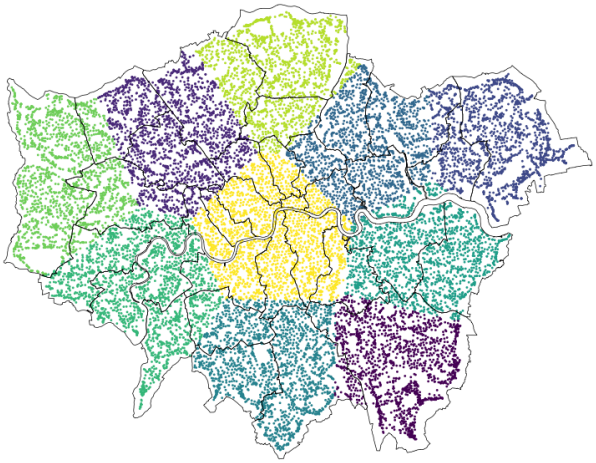

Figure A.1: Map of the spatial k-fold cross-validation splits for hyperparameter tuning and performance assessment (k=10)

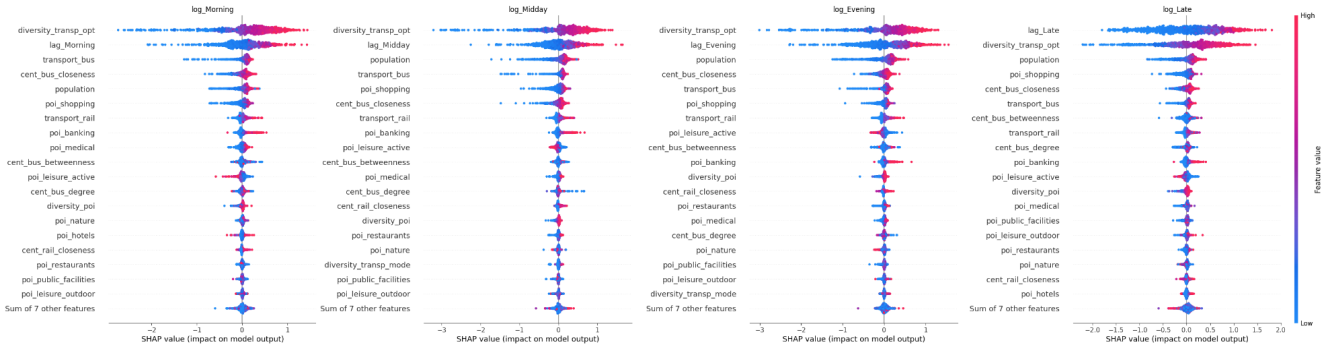

Figure A.2: Summary of feature importance based on SHAP values of a sample. Red-Blue scale represents the feature value, and the x-axis represents the SHAP value, i.e., impact on final model prediction of arrivals in each time band (log)

References

Aaqib Javed, Syed et al. (Apr. 6, 2020). "Estimation of Trip Attraction Rates and Models for Shopping Centers in Dhaka City". In: DOI: 10.5281/zenodo.3733088.

Batty, Michael et al. (Oct. 1, 2014). "Entropy, Complexity, and Spatial Information". In: *Journal of Geographical Systems* 16.4, pp. 363–385. ISSN: 1435-5949. DOI: 10.1007/s10109-014-0202-2. URL: https://doi.org/10.1007/s10109-014-0202-2 (visited on 08/01/2024).

Cervero, Robert and Kara Kockelman (Sept. 1, 1997). "Travel Demand and the 3Ds: Density, Diversity, and Design". In: *Transportation Research Part D: Transport and Environment* 2.3, pp. 199–219. ISSN: 1361-9209. DOI: 10.1016/S1361-9209(97)00009-6. URL: https://www.sciencedirect.com/science/article/pii/S1361920997000096 (visited on 06/14/2024).

Chen, Hugh et al. (June 29, 2020). *True to the Model or True to the Data?* DOI: 10.48550/arXiv.2006.16234. arXiv: 2006.16234 [cs, stat]. URL: http://arxiv.org/abs/2006.16234 (visited on 08/04/2024). Pre-published.

Convery, Sheila and Brendan Williams (Sept. 2019). "Determinants of Transport Mode Choice for Non-Commuting Trips: The Roles of Transport, Land Use and Socio-Demographic Characteristics". In: *Urban Science* 3.3 (3), p. 82. ISSN: 2413-8851. DOI: 10.3390/urbansci3030082. URL: https://www.mdpi.com/2413-8851/3/3/82 (visited on 06/18/2024).

Erlander, Sven and Neil F. Stewart (Dec. 1990). *The Gravity Model in Transportation Analysis: Theory and Extensions*. VSP. 252 pp. ISBN: 978-90-6764-089-3. Google Books: tId3PU1leR8C.

Jayasinghe, Amila, Kazushi Sano, and Kasemsri Rattanaporn (Oct. 1, 2017). "Application for Developing Countries: Estimating Trip Attraction in Urban Zones Based on Centrality". In: *Journal of Traffic and Transportation Engineering (English Edition)* 4.5, pp. 464–476. ISSN: 2095-7564. DOI: 10.1016/j.jtte.2017.05.011. URL: https://www.sciencedirect.com/science/article/pii/S2095756416302458 (visited on 05/17/2024).

Khavarian-Garmsir, Amir Reza et al. (Feb. 2023). "From Garden City to 15-Minute City: A Historical Perspective and Critical Assessment". In: *Land* 12.2 (2), p. 512. ISSN: 2073-445X. DOI: 10.3390/land12020512. URL: https://www.mdpi.com/2073-445X/12/2/512 (visited on 07/29/2024).

Li, Ziqi (Sept. 1, 2022). "Extracting Spatial Effects from Machine Learning Model Using Local Interpretation Method: An Example of SHAP and XGBoost". In: *Computers, Environment and Urban Systems* 96, p. 101845. ISSN: 0198-9715. DOI: 10.1016/j.compenvurbsys.2022.101845. URL: https://www.sciencedirect.com/science/article/pii/S0198971522000898 (visited on 07/07/2024).

Liu, Xiaojian, Ourania Kounadi, and Raul Zurita-Milla (Apr. 2022). "Incorporating Spatial Autocorrelation in Machine Learning Models Using Spatial Lag and Eigenvector Spatial Filtering Features". In: *ISPRS International Journal of Geo-Information* 11.4 (4), p. 242. ISSN: 2220-9964. DOI: 10.3390/ijgi11040242. URL: https://www.mdpi.com/2220-9964/11/4/242 (visited on 07/25/2024).

Louhichi, Mouad et al. (Jan. 1, 2023). "Shapley Values for Explaining the Black Box Nature of Machine Learning Model Clustering". In: *Procedia Computer Science*. The 14th International

Conference on Ambient Systems, Networks and Technologies Networks (ANT) and The 6th International Conference on Emerging Data and Industry 4.0 (EDI40) 220, pp. 806–811. ISSN: 1877-0509. DOI: 10.1016/j.procs.2023.03.107. URL: https://www.sciencedirect.com/science/article/pii/S1877050923006427 (visited on 07/24/2024).

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (visited on 07/24/2024).

MacLeod, Kara E., Brian L. Cole, and Charles Musselwhite (June 2022). "Commuting to Work Post-Pandemic: Opportunities for Health?" In: *Journal of Transport & Health* 25, p. 101381. ISSN: 2214-1405. DOI: 10.1016/j.jth.2022.101381. pmid: 35540370. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9074865/ (visited on 07/24/2024).

Melikov, Pierre et al. (Apr. 1, 2021). "Characterizing Urban Mobility Patterns: A Case Study of Mexico City". In: pp. 153–170. ISBN: 9789811589829. DOI: 10.1007/978-981-15-8983-6_11.

Meyer, Hanna et al. (Nov. 2019). "Importance of Spatial Predictor Variable Selection in Machine Learning Applications – Moving from Data Reproduction to Spatial Prediction". In: *Ecological Modelling* 411, p. 108815. ISSN: 03043800. DOI: 10.1016/j.ecolmodel.2019.108815. arXiv: 1908.07805 [cs, stat]. URL: http://arxiv.org/abs/1908.07805 (visited on 08/01/2024).

Schratz, Patrick et al. (Mar. 29, 2018). *Performance Evaluation and Hyperparameter Tuning of Statistical and Machine-Learning Models Using Spatial Data*.

Shannon, C. E. (July 1948). "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27.3, pp. 379–423. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1948.tb01338.x. URL: https://ieeexplore.ieee.org/document/6773024 (visited on 08/01/2024).

Simini, Filippo et al. (Nov. 12, 2021). "A Deep Gravity Model for Mobility Flows Generation". In: *Nature Communications* 12.1, p. 6576. ISSN: 2041-1723. DOI: 10.1038/s41467-021-26752-4. URL: https://www.nature.com/articles/s41467-021-26752-4 (visited on 05/02/2024).

Travel Forecasting Resource (2024). URL: https://tfresource.org (visited on 05/09/2024).

Wheeler, David and Michael Tiefelsdorf (June 1, 2005). "Multicollinearity and Correlation among Local Regression Coefficients in Geographically Weighted Regression". In: *Journal of Geographical Systems* 7.2, pp. 161–187. ISSN: 1435-5949. DOI: 10.1007/s10109-005-0155-6. URL: https://doi.org/10.1007/s10109-005-0155-6 (visited on 07/25/2024).

Wöhner, Fabienne (June 1, 2022). "Work Flexibly, Travel Less? The Impact of Telework and Flextime on Mobility Behavior in Switzerland". In: *Journal of Transport Geography* 102, p. 103390. ISSN: 0966-6923. DOI: 10.1016/j.jtrangeo.2022.103390. URL: https://www.sciencedirect.com/science/article/pii/S0966692322001132 (visited on 07/24/2024).

Xin, Yanan et al. (Oct. 18, 2022). *Vision Paper: Causal Inference for Interpretable and Robust Machine Learning in Mobility Analysis*. DOI: 10.48550/arXiv.2210.10010.

Yadav, Piyush et al. (June 1, 2017). *The Role of Open Data in Driving Sustainable Mobility in Nine Smart Cities*.

Yang, Yingxiang et al. (July 11, 2014). "Limits of Predictability in Commuting Flows in the Absence of Data for Calibration". In: *Scientific reports* 4, p. 5662. DOI: 10.1038/srep05662.

Zhong, Chen et al. (July 1, 2015). "Measuring Variability of Mobility Patterns from Multiday Smart-Card Data". In: *Journal of Computational Science*. Computational Science at the Gates of Nature 9, pp. 125–130. ISSN: 1877-7503. DOI: 10.1016/j.jocs.2015.04.021. URL: https://www.sciencedirect.com/science/article/pii/S1877750315000599 (visited on 06/18/2024).