Generating Synthetic Energy Usage Data to Enable Machine Learning for Sustainable Accommodation

Peter J. Bentley
Dept. of Computer Science, UCL
AI Lab, Autodesk
London, United Kingdom
p.bentley@cs.ucl.ac.uk

Soo Ling Lim
Dept. of Computer Science
Univerity College London (UCL)
London, United Kingdom
s.lim@cs.ucl.ac.uk

Shrey Jindal
TheSqua.Re
150-152 Fenchurch Street
London, United Kingdom
shrey.jindal@thesqua.re

Sid Narang
TheSqua.Re
150-152 Fenchurch Street
London, United Kingdom
sid.narang@thesqua.re

***Abstract*—Machine Learning has the potential to discover new correlations between energy usage in apartments and variables such as seasonality, apartment location, size, efficiency and details of those staying in the apartments, thus helping apartments to become more sustainable and helping those who stay in them to use less energy. The biggest impedance to creating such ML tools is lack of viable data – without the data, the tools cannot be created – yet it is not feasible to wait for several years' worth of good data before creating the tools. Here we present a solution to this problem: the use of a digital twin to generate synthetic data. This approach is viable even when there is no existing data, but when expert knowledge about the relationship between systems exist. To achieve this, we develop a new agent-based synthetic data generator (ASDG) and explore a case study with a corporate housing and luxury alternate accommodation marketplace called *TheSqua.re*. We show that unlimited quantities of realistic data can be automatically generated, including data for different scenarios, and that it can be used by Machine Learning to discover the underlying correlations.**

***Keywords*—big data, computational modelling, digital twin, synthetic data generator, energy, machine learning, artificial intelligence, sustainability**

I. Introduction

Machine learning (ML) has the potential to transform our ability to understand data and make useful decisions. As we increasingly study energy usage with respect to sustainability, ML tools may help provide solutions to this growing crisis. However, ML algorithms require large quantities of data.

Many applications of machine learning (ML) in industry settings often face the problem of not having access to sufficient data to train and test their ML algorithms. There are many reasons this can happen: (1) the existing dataset has missing or inaccurate data, as the existing data collection process was not conducted for the purpose of ML, (2) the data is sensitive and protected by data protection laws (frequently the case with data relating to people), (3) the development of the ML model for a given application is at an early stage and the data for that application does not yet exist or is insufficient in quality or quantity to perform ML training, (4) the data of interest comprises rarely occurring edge-cases or anomalies, (5) the real data is skewed or not representative, (6) data collection at the scale required to train the ML model is too costly or time consuming, and (7) it may be difficult or infeasible to label the data, making it more difficult to use supervised ML approaches.

In this work we focus on the problem of improving systems relating to sustainable rental accommodation. With most organisations now careful to minimize their carbon footprints, there is a new requirement to find environmentally sound accommodation options for essential business travel. One side effect of the massive recent disruption caused by the global pandemic appears to be a change of behaviour – business travellers are now travelling less, but staying longer for each trip. This increases the significance of their accommodation: the more business travellers who stay in energy-efficient accommodation (including access to green transportation) the less environmental impact their business trip will have.

Rented apartments are a recognized sustainable alternative to hotels [1]. Such apartments can be analysed and have improvements made to enhance insulation and reduce energy consumption further. Appliances could be optimized to reduce energy. Habits of frequent guests could be tracked so that the needs of each guest are matched to the type of apartment (i.e., those who prefer high internal temperatures should be given apartments with more efficient insulation and heating). ML can be used to learn the operational characteristics of apartments with different guests in order to achieve these goals.

This work focusses on the problem of providing these solutions for business apartments. We provide a real case study: a digital corporate housing and luxury alternate accommodation marketplace called *TheSqua.re* with 200,000 furnished apartments managed by 2000 plus operators globally. Despite being a market leader in data collection and analysis for this domain, all of the data issues listed previously exist, and all for unavoidable reasons. (1) There is little usable data on apartment energy usage as it was not possible to read meters during the pandemic, and difficult to train staff to submit perfectly accurate readings every month. (2) Data relating to guests is private and cannot be shared. (3) the use of ML is new for this organization and appropriate data collection has only been ongoing for 12 months. (4) Some data of interest may involve rare edge cases, such as misuse of apartments by guests or staff. (5) Because of the recent pandemic which has severely altered recent patterns of travel, all recent data is skewed and not representative of normal behaviour. (6) Perfect data collection of energy usage and actual guest behaviours within apartments is costly. While

978-1-6654-1262-9/21/$31.00 ©2021 IEEE

smart meters and other smart energy monitoring devices are being installed over the coming years, the process is too slow to enable sufficient data in order to generate the ML solutions now. (7) The data is difficult to label: was the energy consumption caused by guest A or guest B? Were they children or adults? Was it produced by heating or cooking? Was the high energy usage for an empty apartment caused because the heating was left on by mistake or because it is a false meter reading?

With useful data only just starting to be created and with the effects of the pandemic far from over, this severely impacts the feasibility of creating ML solutions to the problem of providing more sustainable options for business travellers (a problem faced not just in this industry but throughout all industries attempting to tackle sustainability). We need a new source of data to enable the appropriate ML systems to be built now so that they will be ready for use when the data is also ready.

We propose a method that can generate synthetic human behavior data for ML models. This method uses computational modelling to act as a digital twin of a real marketplace, producing realistic, and pre-labelled datasets for ML models. The method enables the generation of large quantities of data at scale, and the ability to incorporate different scenarios, noise and bias, so that it can be used to test the ability of ML to learn behaviours. As the data that is generated is completely synthetic, it does not contain sensitive information or personally identifiable information, hence is safe to use (as opposed to anonymized data, which runs the risk of being re-identified [2]).

The paper makes the following contributions: (1) we proposed an agent-based synthetic data generator (ASDG) to generate synthetic energy usage data, (2) we applied the ASDG to generate synthetic data in a real-world case study of a short-term rental apartment marketplace, and (3) we demonstrated the application of ML on the synthetic data.

II. Background

A. Data Augmentation

Data augmentation is the process of enhancing the size and quality of datasets by adding slightly modified copies of the existing dataset. Data augmentation has been used extensively in computer vision ML [3]. Some commonly used approaches for image classification include zooming in, cropping, rotating, flipping or distorting the original image [4], and erasing a random region in the image with random values [5]. Data augmentation is shown to be an effective method to improve ML and reduce overfitting, however, it only be used when the data already exists. In addition, the quality of the augmented dataset depends on the quality of the original dataset.

B. Synthethic Data Generators (SDGs)

Synthetic data generators create artificial data by using different algorithms that mirror the properties of the original data. Similar to data augmentation, many SDGs depend on the existence of datasets in order to work. ML methods such as generative adversarial networks (GANs) can be used to generate synthetic data [6]. For example, Fekri et al. [7] developed a recurrent generative adversarial network (R-GAN) for generating realistic energy consumption data by learning from existing smart grid data. Jälkö et al. [8] uses probabilistic modelling to produce strongly anonymized synthetic data and showed that the same statistical discovery could be made from the synthetic data as with the original data, and the usability of the synthetic data depends on the quality of the models. Amodio et al. [9] developed a model to generate hard-to-obtain information from easy-to-obtain information and showed that their finding has applications in drug discovery and clinical inference. Mapping between the easy-to-collect and hard-to collect information can be trained as a conditional GAN from a subset of samples with both measured. With their conditional GAN model known as feature mapping GAN (FMGAN), the results of expensive experiments can be predicted, saving on the costs of actually performing the experiment.

C. Computational Modelling

Chen and Venkatachalam [10] proposed a process-based definition of big data, where big data should be perceived as a continuous, unstructured and unprocessed dynamics of primitives, rather than as points (snapshots) or summaries (aggregates) of an underlying phenomenon. According to them, agent-based models often generate big data, but researchers tend not to store, retrieve and reuse the data, partly because of the associated demands on memory, but also due to the relative lack of awareness concerning the use of ML to handle it [10]. Marketplace models exist (e.g., [11, 12]), but they are mainly used to understand marketplaces rather than generate data. Richardson et al. [13] developed an agent-based model for domestic electricity use, but it is not specifically for ML and their model takes no account of the attitudes of occupants toward energy use. Researchers have also started to use computational modelling to generate training data for ML in animal behavior and shown that synthetic data can help improve clustering [14]. There may also be potential for the emerging field of agent-based modelling of human behaviour to be used for data generation [15, 16].

III. Method

In this problem there is insufficient existing data for synthetic data generators using methods such as generative adversarial networks (GANs), or variational autoencoders (VAEs) and the complexity of interactions mean that statistical models are unlikely to capture causal effects correctly. However, considerable high-level knowledge exists about the nature of the marketplace, its constituents, and the interactions between them. This means that the most suitable form of SDG in this case is through agent-based modelling. Here we propose a model that can generate synthetic data for the problem of ML to enable sustainable business accommodation, using agent-based modelling: an agent-based synthetic data generator (ASDG). Fig. 1 shows the main components of the ASDG. Fig. 2 provides the algorithm. We then describe the ASDG in more detail with Table I providing the values for model constants.

A. Initialise Marketplace

Create N_{build} buildings, where each building has:

- A randomly selected location, $B_{loc} = loc_{random} \in \mathbb{Z}: [1..N_{loc}]$.
- A randomly selected community heating setting, $B_{community} = community_{random} \in \mathbb{Z}: [0,1]$, where 1

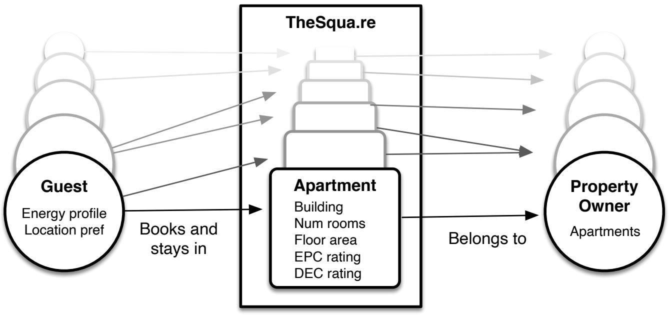

Fig. 1. The components of the marketplace and their attributes. The marketplace displays apartments to guests and matches their booking requests with available apartments.

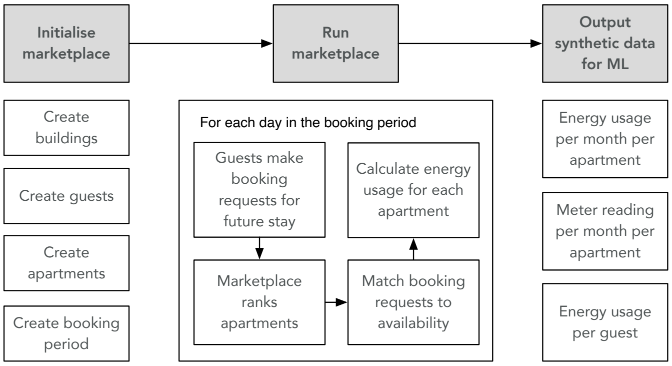
Initialise marketplace
Run marketplace
Output synthetic data for ML
Create buildings
Create guests
Create apartments
Create booking period
For each day in the booking period
Guests make booking requests for future stay
Calculate energy usage for each apartment
Marketplace ranks apartments
Match booking requests to availability
Energy usage per month per apartment
Meter reading per month per apartment
Energy usage per guest

Fig. 2. The ASDG algorithm.

means the building provides community heating and hot water for all apartments (a more efficient method compared to individual boilers within each apartment) and 0 means it does not.

Create N_{apt} apartments, where each apartment has:

- A randomly selected building that it is located in, $A_{build} = b_{random} \in \mathbb{Z}: [1..N_{build}]$.
- A randomly selected number of bedrooms that it contains, $A_{rm} = r_{random} \in \mathbb{Z}: [N_{rm\_min}..N_{rm\_max}]$.
- Floor area (in m^2), $A_{area} = a_{common} + \sum_{i=1}^{Arm} a_i$, where a_{common} is a randomly selected floor area for common areas such as kitchen, bathroom and sitting room, $a_{common} = a_{random} \in \mathbb{Z}: [a_{common\_min}..a_{common\_max}]$ and for each room i in the apartment, a_i is a randomly selected floor area where $a_i = a_{random} \in \mathbb{Z}: [a_{rm\_min}..a_{rm\_max}]$.
- A random energy performance certificate (EPC) rating, $A_{epc} = EPC_{rand} \in \{A, \dots, E\}$, with probability $P_1(EPC_{rand})$ if $B_{community} = 1$, or $P_2(EPC_{rand})$ otherwise. Table I provides the probabilities. (A property needs to have an EPC of at least E to be rentable. Apartments with community heating can normally only have EPC ratings of A, B or C.)
- A display energy certificate (DEC) rating, which represents operational usage, where $A_{dec} = A_{epc}$ with $P(0.7)$, otherwise $A_{dec} = EPC_{rand} \in \{A, \dots, E\}$ with equal probability of selecting any rating. This models the fact that the operation of the apartment can sometimes result in significantly different energy consumption compared to that predicted by the EPC, which may be caused by new appliances or changes made since the EPC was produced.

TABLE I. MODEL CONSTANTS AND VALUES

Constant	Value	Constant	Value	Constant	Value
N_{build}	10	$P_1(EPC_{rand} = A)$	0.4	N_{booker}	100
N_{loc}	8	$P_1(EPC_{rand} = B)$	0.4	N_{year}	5
N_{apt}	50	$P_1(EPC_{rand} = C)$	0.2	$N_{book\_gap\_min}$	7
N_{rm_min}	1	$P_1(EPC_{rand} = D)$	0.0	$N_{book\_gap\_max}$	180
N_{rm_max}	2	$P_1(EPC_{rand} = E)$	0.0	$N_{stay\_gap\_min}$	21
a_{common_min}	30	$P_2(EPC_{rand} = A)$	0.0	$N_{stay\_gap\_max}$	180
a_{common_max}	40	$P_2(EPC_{rand} = B)$	0.0	P_{d1}	0.3
a_{rm_min}	10	$P_2(EPC_{rand} = C)$	0.6	P_{d3}	0.1
a_{rm_max}	15	$P_2(EPC_{rand} = D)$	0.3	P_{d7}	0.4
		$P_2(EPC_{rand} = E)$	0.1	P_{d14}	0.1
		N_{pot_guest}	1000	P_{d21}	0.1

Create N_{pot_guest} potential guests, where each guest has:

- A randomly selected energy usage level, $G_{energy} = energy_{random} \in \{0,1,2\}$, where 0 means the guest has very low energy usage, 1 is medium, 2 is high.
- A randomly ordered location preference, G_{loc_pref} =random order of $\{1, \dots, N_{loc}\}$.

Create booking period with the following attributes:

- A start date, $start\_date_{actual}$ = 1st January 2021
- An end date, $end\_date_{actual} = start\_date_{actual} + N_{year}$
- A simulated start date, $start\_date_{sim} = start\_date_{actual} - 2\ months$. This is to run the simulation for a few months before actual start date in order to warm up the marketplace, (e.g., get the bookings to a realistic level).
- A simulated end date, $end\_date_{sim} = end\_date_{actual} + 2\ months$. This is to avoid any artefacts caused by a forced end to the booking availability.

The booking period determines the simulation duration, beginning from $start\_date_{sim}$ and ending at end_date_{sim}.

B. Run Marketplace

For each day d in the booking period, a random set of N_{booker} guests out of N_{pot_guest} are chosen. For each guest in the set, we create a booking request with the following attributes:

- A start date for the stay, $start\_date_{stay}$. If this is the guest's first booking, then a random date is chosen, $start\_date_{stay} = d + N_{random} \in \mathbb{Z}: [N_{book\_gap\_min} .. N_{book\_gap\_max}]$. This is to simulate the guest booking for a stay between 1 week and 6 months in the future. If the guest has an existing booking, $start\_date_{stay} = end\_date_{most\_recent\_stay}$ + $N_{random} \in \mathbb{Z}: [N_{stay\_gap\_min} .. N_{stay\_gap\_max}]$, where $end\_date_{most\_recent\_stay}$ is the end date of the most recent stay. We assume that the guest always books the next trip in the future of the most recent trip with a reasonable gap in between.

- An end date for the stay, end_date_{stay}=$start\_date_{stay} + dur_{stay} - 1$, where $dur_{stay} = dur_{rand} \in \{1,3,7,14,21\}$, where $P(dur_{rand} = 1) = P_{d1}, P(dur_{rand} = 3) = P_{d3}, \ldots, P(dur_{rand} = 21) = P_{d21}$, where $P_{d1} + P_{d3} + \cdots + P_{d21} = 1$. The duration simulates the typical duration for business travels.

- A random number of guests for the booking, $N_{bguest} = N_{random} \in \mathbb{Z}: [N_{rm\_min} .. N_{rm\_max}]$. This assumes each bedroom of an apartment is occupied by at most one guest, and no more, for simplicity.

In response to each booking request, the marketplace presents the available apartments in a ranked order to the guest. There are two options for the marketplace ranking:

1. Location ranking (LocRank), where apartments are ranked by the guest's preferred location

2. Eco-friendly ranking (EcoRank), where apartments are ranked by guest's preferred location, and then by its EPC rating.

For each apartment in the ranked list of apartments, if the highest ranked apartment has attributes that match the requested number of guests and duration of the booking, it is booked, otherwise, the next apartment in the list is checked. This simulates the guest looking on the platform to find apartments that match their booking request. For now, we assume booking requests are fixed – so guests do not change their booking requests to match availability, which is a possible scenario. If nothing matches the booking request, it fails and no booking is made (this simulates the common purchase behaviour of people trying other alternatives).

Once all bookings have been processed for day *d*, the model then calculates the energy usage of each apartment *A* for day *d*. If *A* is occupied (there is a booking for the day), the energy usage for A, $A_{energy\_usage} = E_{est} \times T_{effect} \times G_{effect} \times P_{effect} + rand(0{,}5)$, where:

Energy estimate $E_{est} = E_{estsqmy} \times floor\_area/365$, where $E_{estsqmy}$ is the per square meter yearly energy estimate of the apartment based on actual EPC, $E_{estsqmy} = E_{rand(EPC)} \in \mathbb{Z}: [E_{\min(EPC)} .. E_{\max(EPC)}]$ (Table II).

Temperature effect, T_{effect}, if $T_d < 10°\text{C}$, $T_{effect} = 1$, if T_d is in 11-15°C, $T_{effect} = 0.8$, else $T_{effect} = 0.3$, where $T_d = T_{rand} \in \mathbb{Z}: [T_{min} .. T_{max}] + T_{var}$, with the values of $T_{min}$and T_{max} determined by the specific month the current day d falls within (Table III), and random variation $T_{var} = T_{rand} \in \mathbb{Z}: [T_{\min\_var} .. T_{\max\_var}]$, where $T_{\min\_var} = 0, T_{\max\_var} = 5$.

Number of guest effect, G_{effect}, if $N_{guest} = 1$, $G_{effect} = 1$, if $N_{guest} = 2$, $G_{effect} = 1.5$, else $G_{effect} = 1.8$

Guest energy profile effect, P_{effect}, if energy usage of guest $G_{energy} = 0$, then $P_{effect} = 0.5$, $G_{energy} = 1$, then $P_{effect} = 1$ else $P_{effect} = 1.5$

If *A* is not occupied, there is a low-level energy usage. The energy usage for A, $A_{energy\_usage} = E_{est} \times T_{effect} \times P_{effect\ (Eguest=0)} + rand(0.5{,}1.5)$, where $E_{est} = E_{rand} \in \{11{,}2{,}3{,}4{,}5{,}6{,}7{,}12\}$, where $P(E_{rand} = 1) = \frac{5}{30}, P(E_{rand} = 2) = \frac{5}{30}, P(E_{rand} = 3) = \frac{5}{30}, P(E_{rand} = 4) = \frac{5}{30}, P(E_{rand} = 5) = \frac{5}{30}, P(E_{rand} = 6) = \frac{2}{30}, P(E_{rand} = 7) = \frac{2}{30}, P(E_{rand} = 12) = \frac{1}{30}$.

C. *Output Synthetic Data for Machine Learning*

For every day in the simulation between $start\_date_{actual}$ and end_date_{actual}, and for each apartment, the system outputs the following synthetic data for apartment daily energy usage. Once the simulation completes, ASGD further groups the output data by month to produce, for each apartment, monthly energy usage, monthly meter reading and for each guest their energy usage, see Fig. 3.

A run of the simulation takes approximately 1 second CPU time per simulated year on a MacBook Air (M1, 2020) with 8 cores and 16GB memory. The simulation is implemented in Python and the code can be requested from the authors via email.

IV. Case Study

We applied the model on a case study of an accommodation marketplace: TheSqua.re. We aimed to make a digital twin of this marketplace by calibrating all model constants and values to match existing data and knowledge from TheSqua.re.

TABLE II. Yearly Min and Max Estimated Energy for each EPC

Energy (kWhm^2)	**EPC**				
	A	***B***	***C***	***D***	***E***
$E_{\min\ (EPC)}$	10	15	70	180	350
$E_{\max\ (EPC)}$	100	90	240	440	600

TABLE III. Month and Temperature[1]

Temp (°C)	**Month (Jan-Dec)**											
	Jan	***Feb***	***Mar***	***Apr***	***May***	***Jun***	***Jul***	***Aug***	***Sep***	***Oct***	***Nov***	***Dec***
T_{min}	3	3	4	6	9	12	14	14	11	9	6	3
T_{max}	8	9	12	16	18	21	24	22	20	16	12	9

[1] Temperature information from https://www.timeanddate.com/weather/uk/london/climate

Apartment Daily Energy Usage

Date
Temperature
Apartment ID
Building ID
Number of rooms
Floor area
EPC rating
On community heating? (Y/N)
Is occupied? (Y/N)
Number of guests
Guest energy usage profile (or empty if the apartment is vacant on the day)
Energy usage

Apartment Monthly Energy Usage

Month and year
Average temperature
Apartment ID
Building ID
Number of rooms
Floor area
EPC rating
On community heating? (Y/N)
Total number of days occupied
Average number of guests per day
Average guest energy usage profile
Total energy usage for the month

Apartment Monthly Meter Reading

Month and year
Apartment ID
Building ID
Number of rooms
Floor area
EPC rating
On community heating? (Y/N)
Total number of days occupied
Average number of guests per day
Meter reading (Cumulative energy usage)
Primary energy use (Monthly energy usage x 12 / Floor area)

Guest Energy Usage

Guest ID
Energy usage profile
Total number of stays
Total number of days stayed
Average number of days per stay
Total energy usage per stay
Average energy usage per day

Fig. 3. Synthetic datasets generated by the ASDG, produced per apartment, per day/month, per guest.

TheSqua.re is a digital corporate housing and luxury alternate accommodation marketplace with 200,000 furnished apartments managed by 2000 plus operators in 600 Cities globally. They provide a large choice of alternate accommodations in major cities, using proprietary technology which includes a customisable booking platform designed to allow enterprise customers manage Carbon Neutral Global travel programmes. In 2020, they launched MySqua.re, a private label brand that delivers a portfolio of homes in London offering city centre locations, currently live in more than ten neighbourhoods in London such as Fitzrovia, Mayfair, Kensington, Canary Wharf and City of London, operating more than 100 apartments. Our case study focusses on these London-based properties.

TheSqua.re Group aims to reach net zero by 2023. They are doing this by analysing data including accommodation EPC, proximity of accommodation to shared and electric mobility, sources of renewable energy used in the apartments, and recycling availability in the buildings. Through their API they will power other vendors of temporary accommodations so they can offer low carbon housing to their corporate and luxury clients worldwide. To achieve this goal, it is necessary to enhance understanding of energy usage within apartments. Currently, because of the reasons listed in the introduction, it is not able to predict a fair energy usage policy for the apartments. It is anticipated that there will be a correlation between apartment usage and factors such as seasonality, origin country/location of the guest, apartment location, size, and efficiency. However, the correlation is likely to be nonlinear and require ML to discover it. Without data, the ML cannot be built, hence in this case study we generate synthetic data to enable ML prediction of monthly apartment energy usage.

A. Experiments

We run our agent-based synthetic data generator to generate data for both ranking methods, EcoRank and LocRank, each for the duration of 1, 2, 3 and 4 years, which resulted in 8 datasets in total. For the experiments, $N_{apt} = 50$, making the size of the 1-year dataset N=600, 2-year dataset N=1200, 3-year dataset N=1800, 4-year dataset N=2400.

To demonstrate how the generated data can be used for ML, we used two ML algorithms, Linear Regression and Multi-layer Perceptron Regressor (MLPRegressor). We used scikit-learn (https://scikit-learn.org/) and the 'lbfgs' solver setting for MLPRegressor, which is an optimiser in the family of quasi-Newton methods. The features presented for learning were: average temperature (temperature), number of rooms, EPC rating (EPC), floor area, community heating (community), occupancy, average number of guests per day (num guests), average guest energy usage profile (guest profile), in order to predict apartment energy usage. We used 10-fold cross validation.

B. Results

The results are shown in Fig. 4. The results confirm that there is a nonlinear correlation between input variables and energy usage in the apartments as the accuracy for Linear Regression is consistently lower (0.09-0.52) compared to accuracy for MLP Regressor (0.51-0.91). Overall, for both methods, the accuracy is the lowest for the 1-year dataset, and continues to improve (either with higher accuracy, or lower standard deviation, or both) as the number of years increases in the dataset. MLP Regressor performs better in EcoRank compared to LocRank, suggesting that the more guests are given the opportunity to choose sustainable apartments, the more predictable the energy usage of the apartments becomes.

To understand the importance of the features in their use by the ML model when making predictions, we ran feature permutation to see which features are more important. Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target

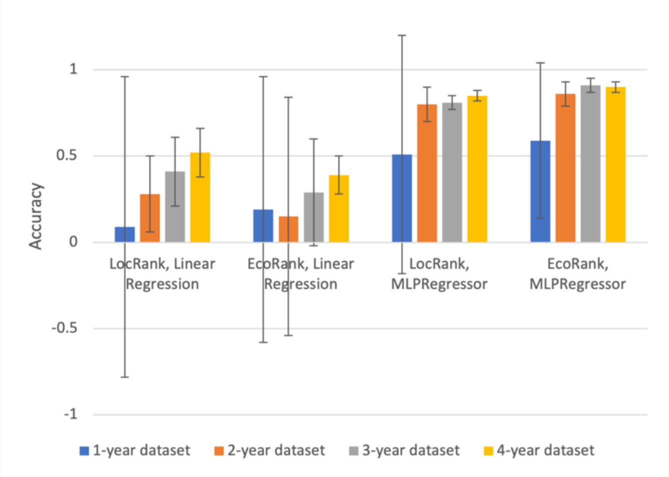

Fig. 4. Accuracy of Linear Regression and MLPRegressor on all synthetic datasets when predicting energy usage. Error bars: mean ± standard deviation.

variable. Overall, the top three features found by both ML methods are: temperature, EPC and guest profile, showing that they successfully identify the key factors in the model, despite the large amount of random noise and variability (Table IV).

V. Conclusion and Future Work

This paper proposed a new agent-based synthetic data generator (ASDG), which generates synthetic data in order to enable ML tools to be generated now that will reduce energy usage in short-term rental apartments. The ASDG overcomes the problem of having insufficient existing data for synthetic data generators using methods such GANs or VAEs, and instead uses knowledge about the nature of the marketplace, its constituents, and the interactions between them in order to behave as a digital twin of a real corporate housing and luxury alternate accommodation marketplace. The result is a scalable SDG that can generate a year’s worth of labelled data per second, for multiple alternative scenarios (including those that may not yet exist) which we have shown can be used by ML algorithms to learn the correlations within the data.

TABLE IV. Top 3 Features for each Dataset, with Corresponding Average Permutation Importance and Standard Deviation in Brackets

Feature Ranking	**Dataset (LocRank, Linear Regression)**			
	1-year	***2-year***	***3-year***	***4-year***
Rank 1	Temperature 0.300 (0.051)	Temperature 0.319 (0.043)	EPC 0.721 (0.055)	Temperature 0.512 (0.035)
Rank 2	EPC 0.259 (0.061)	EPC 0.175 (0.028)	Temperature 0.526 (0.043)	EPC 0.510 (0.035)
Rank 3	Num rooms 0.156 (0.032)	Community 0.116 (0.025)	Num rooms 0.125 (0.018)	Guest profile 0.092 (0.015)

Feature Ranking	**Dataset (EcoRank, Linear Regression)**			
	1-year	***2-year***	***3-year***	***4-year***
Rank 1	EPC 0.518 (0.081)	Temperature 0.593 (0.052)	Temperature 0.318 (0.034)	Temperature 0.381 (0.026)
Rank 2	Temperature 0.375 (0.056)	EPC 0.523 (0.071)	Community 0.289 (0.031)	EPC 0.247 (0.023)
Rank 3	Floor area 0.195 (0.033)	Guest profile 0.112 (0.026)	Num rooms 0.069 (0.013)	Num guests 0.143 (0.016)

Feature Ranking	**Dataset (LocRank, MLPRegressor)**			
	1-year	***2-year***	***3-year***	***4-year***
Rank 1	EPC 0.565 (0.105)	Temperature 0.418 (0.067)	Temperature 0.576 (0.077)	Temperature 0.698 (0.041)
Rank 2	Temperature 0.565 (0.112)	EPC 0.390 (0.064)	EPC 0.563 (0.062)	EPC 0.517 (0.051)
Rank 3	Guest profile 0.200 (0.062)	Guest profile 0.159 (0.026)	Guest profile 0.154 (0.020)	Guest profile 0.153 (0.020)

Feature Ranking	**Dataset (EcoRank, MLPRegressor)**			
	1-year	***2-year***	***3-year***	***4-year***
Rank 1	EPC 1.209 (0.297)	EPC 0.952 (0.150)	Temperature 0.461 (0.048)	EPC 0.605 (0.105)
Rank 2	Temperature 0.866 (0.138)	Temperature 0.783 (0.101)	Community 0.346 (0.031)	Temperature 0.500 (0.042)
Rank 3	Guest profile 0.161 (0.057)	Num guests 0.194 (0.032)	EPC 0.143 (0.031)	Floor area 0.209 (0.057)

Our experiments illustrate the success of this approach and demonstrate that the best results from ML require two or more years data – meaning the use of synthetic data can enable the creation of ML for sustainable applications significantly faster than otherwise possible, given the difficulties still faced in obtaining data for such domains.

Future work will add pricing to the model, other reasons for travel (e.g., leisure) and multiple scenarios to enable A/B testing for marketplace settings. We also anticipate that a more generic marketplace model can be created that can then be applied to the creation of synthetic data to other marketplaces.

References

[1] Y. V. Palgan, L. Zvolska, and O. Mont, "Sustainability framings of accommodation sharing," *Environmental Innovation and Societal Transitions,* vol. 23, pp. 70-83, 2017.

[2] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A systematic review of re-identification attacks on health data," *PloS ONE,* vol. 6, p. e28071, 2011.

[3] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data,* vol. 6, pp. 1-48, 2019.

[4] M. D. Bloice, C. Stocker, and A. Holzinger, "Augmentor: an image augmentation library for machine learning," *arXiv:1708.04680,* 2017.

[5] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 13001-13008.

[6] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621,* 2017.

[7] M. N. Fekri, A. M. Ghosh, and K. Grolinger, "Generating energy data for machine learning with recurrent generative adversarial networks," *Energies,* vol. 13, p. 130, 2020.

[8] J. Jälkö, E. Lagerspetz, J. Haukka, S. Tarkoma, A. Honkela, and S. Kaski, "Privacy-preserving data sharing via probabilistic modeling," *Patterns,* vol. 2, p. 100271, 2021.

[9] M. Amodio, D. Shung, D. Burkhardt, P. Wong, M. Simonov, Y. Yamamoto, D. van Dijk, F. P. Wilson, A. Iwasaki, and S. Krishnaswamy, "Generating hard-to-obtain information from easy-to-obtain information: applications in drug discovery and clinical inference," *Patterns,* vol. 2, p. 100288, 2021.

[10] S.-H. Chen and R. Venkatachalam, "Agent-based modelling as a foundation for big data," *Journal of Economic Methodology,* vol. 24, pp. 362-383, 2017.

[11] S. L. Lim and P. J. Bentley, "Investigating app store ranking algorithms using a simulation of mobile app ecosystems," in *Proceedings of the IEEE Congress on Evolutionary Computation*, 2013, pp. 2672-2679.

[12] S. L. Lim, P. J. Bentley, and F. Ishikawa, "The effects of developer dynamics on fitness in an evolutionary ecosystem model of the App Store," *IEEE Transactions on Evolutionary Computation,* vol. 20, pp. 529-545, 2015.

[13] I. Richardson, M. Thomson, D. Infield, and C. Clifford, "Domestic electricity use: a high-resolution energy demand model," *Energy and Buildings,* vol. 42, pp. 1878-1887, 2010.

[14] L. A. Bolaños, D. Xiao, N. L. Ford, J. M. LeDue, P. K. Gupta, C. Doebeli, H. Hu, H. Rhodin, and T. H. Murphy, "A three-dimensional virtual mouse generates synthetic training data for behavioral analysis," *Nature Methods,* vol. 18, pp. 378-381, 2021.

[15] S. L. Lim and P. J. Bentley, "Coping with uncertainty: modelling personality when collaborating on noisy problems," in *Proceedings of the Artificial Life Conference*, 2018, pp. 566-573.

[16] S. Guo, S. L. Lim, and P. J. Bentley, "Teams frightened of failure fail more: modelling reward sensitivity in teamwork," in *Proceedings of the IEEE Symposium Series on Computational Intelligence*, 2020, pp. 109-116.