Business Understanding

The BCA Green Mark scheme was launched in January 2005 as an initiative to drive Singapore's construction industry towards more environment-friendly buildings. (2023 Building and Construction Authority). Given that the scheme has been in existence for close to two decades, it is timely to review the effectiveness of the scheme. The BCA Green Mark scheme is targeted at developers, building owners and government agencies for new and existing buildings, districts, parks, infrastructure and building interiors. The five key criteria used by the BCA to accord Green Mark statuses are (1) climatic responsive design; (2) building energy consumption; (3) resources stewardship; (4) smart and healthy building; and (5) advance green efforts. Each Green Mark certification is valid for three years from the date of Letter of Green Mark Award issued by BCA. Additionally, there are various Green Mark incentive schemes in place through cash or gross floor area incentives. In this project, the dataset is with a focus on building owners, buildings and their 2017 and 2018 yearly energy consumption.

The business goal is to validate the effectiveness of the BCA Green Mark Scheme with the corresponding data mining goal to identify if certain types of buildings are more (or less) likely to have low yearly energy consumption through the identification of correlation between buildings' profile (e.g. building type, green mark rating) and their energy consumption. This goal will be achieved using the Apriori Algorithm method. Apriori Algorithm is identified to determine the buildings' profile frequently occurring with high probability.

Data Understanding

The dataset on building energy performance data collected through BCA's Building Energy Submission under the legislation on Annual Mandatory Submission of Building Information and Energy Consumption Data for Section 22FJ 'Powers to Obtain Information' of Building Control Act will be utilized for the analysis. (https://data.gov.sg/dataset/building-energy-performance-data).

The following is a summary of the variables recorded and their associated data types, purpose and data quality issues:

Attribute Name	Data Types	Purpose	Data Quality Issues
buildingname	Nominal	Unique identification of building's location	Some missing entries indicated with "NA"
buildingaddress	Nominal	Unique identification of building's location	Some missing entries indicated with "NA"
buildingtype	Nominal	Type of building based on usage of the building. The types are: Community hospital, General hospital/specialist centre (public), Hotel, ITE, Mixed development, Nursing home, Office, Polyclinic, Polytechnic, Private clinic, Private	No issues

		college, Private hospital (private), Private school, Retail, Specialist Centre (public), TCM clinic, University.	
greenmarkstatus	Binary	Indication if the building has Green Mark certification. Yes: Green Mark building. No: Non-Green Mark building	No issues
greenmarkrating	Ordinal	Type of Green Mark rating obtained (if any) Green Mark Ratings are: Certified, Gold, GoldPlus, Legislated, Platinum and NA.	Entries with "NA" could be either due to missing entries or the building is Non-Green building.
greenmarkyearaward	Discrete	Year that the building obtained Green Mark certification (if any). For Non-Green Mark buildings, "NA" is indicated.	Entries with "NA" could be either due to missing entries or the building is Non-Green building.

grossfloorarea	Discrete	Gross floor area of the building	Some missing entries indicated with “NA”
2017energyuseintensity	Discrete	Energy consumption in Year 2017	Some missing entries indicated with “NA”
2018energyusintensity	Discrete	Energy consumption in Year 2018	Some missing entries indicated with “NA”
voluntarydisclosure	Binary	Indication of the data being provided on voluntary basis. Y: Yes N: No	No issues.

The following is an example of sample records with data values for the first six records.

buildingname	buildingaddress	buildingtype	greenmarkstatus	greenmarkrating	greenmarkyearaward	grossfloorarea	2017energyuseintensity	2018energyusintensity	voluntarydisclosure
HEXACUBE	160 CHANGI ROAD, SINGAPORE 419728	Mixed Development	No	NA	NA	5,036	81	105	Y
NA	NA	Retail	No	NA	NA	NA	475	402	N
CITY SQUARE MALL	180 KITCHENER ROAD, SINGAPORE 208539	Retail	Yes	Platinum	2018	65,640	382	365	Y
REPUBLIC PLAZA	9 RAFFLES PLACE, SINGAPORE 048619	Office	Yes	Platinum	2018	102,356	212	183	Y
CENTRAL MALL	1 MAGAZINE ROAD, SINGAPORE 059567	Office	Yes	Platinum	2017	15,769	203	181	Y
KINEX	11 TANJONG KATONG ROAD, SINGAPORE 437157	Retail	Yes	Gold	2015	28,938	431	396	Y

Data preparation

Two data quality issues were identified in the data understanding stage; missing entries and entries with "NA" records could be either due to missing entries or the building is classified without a "greenmarkrating". Out of the 1,244 recorded data, 258 records that were obtained through non-voluntary disclosures consistently had variables such as building names, building address, green mark rating, year of award of green mark and gross floor area indicated by a "NA" for each variable.

Given that these incomplete records constitutes roughly 21% of the total records and aggregation is unable to impute a reasonable data based on average or using buildings with similar profile, the usage of such records with missing entries will skew the data mining performance and outcome of the solution. Hence, removal of these missing records from further analysis will assist with a clearer output. With the removal of data obtained through non voluntary disclosure, the issue of being unable to distinguish whether an "NA" entry is attributable to a missing entry or is a Non-Green Mark building in the columns "greenmarkrating " and "greenmarkyearaward" will consequently be eliminated.

The variables "buildingname" and "buildingaddress" are unique identifiers of the buildings. As these variable are potentially not useful for use in finding correlations, we will not be using these variables in the Apriori Algorithm.

There are two variables (i.e. "2017energyuseintensity" and "2018energyusintensity") in the dataset that provide the energy consumption of the building, each reflecting the energy consumption recorded for the respective buildings in 2017 and 2018. Since the aim is to generate rules involving the attributes related to the building with energy consumption, there should only be a single attribute relating to energy consumption. A new attribute will be derived based on the average of energy consumption in 2017 and 2018.

Additionally, "https://www.archtoolbox.com/energy-use-intensity/" which states that "Simply measuring the amount of energy used per a chosen time period does not take into account

building size, configuration or type of use. The use of Energy Use Intensity (EUI) indicator provides the means to equalise the way that energy use is compared between various types of buildings, and evaluate the means of reducing overall energy consumption.". Based on this reference, the use of EUI would be a more accurate comparison for the different building types. As such, we will convert the newly derived attribute of average energy consumption to EUI using the following formula:

EUI = average energy consumption for 2017 and 2018 / gross floor area
With this conversion, the variables "grossfloorarea", "2017energyuseintensity" and "2018energyusintensity" will collectively be represented by "EUI".

As Apriori algorithm can only use categorical data, we will be transforming the following numeric variables. The variable "Greenmarkyearaward" is a discrete variable and we would convert it to "Nominal" in the SPSS modeler. We will apply the binning method to the variable "EUI". The choice of binnings is derived based on the computation of the minimum, median and maximum data for the "EUI" variable as follows:

	EUI
Minimum	2.13E-05
Median	0.022192
Maximum	4.860714

The following binnings will be used.

	Bin 1	Bin 2	Bin 3
EUI	<0.01 (low)	0.01 < x < 0.5 (middle)	>0.5 (high)

Modelling

The analysis goal will be achieved using the Apriori algorithm method. Apriori algorithm is identified to determine the buildings' profile frequently occurring with high probability, leveraging on rules generated that have correlation with energy consumption. Given that energy consumption is one of the criteria for the award of BCA Green Mark certification, with the knowledge of the correlations with energy consumption (as represented by “EUI” attribute after the data preparation stage), the BCA policy holder can adjust the policy in aspects relating to the target audience for outreach, adjustment that can be made to the incentive scheme for certain groups, depending on the list of the best-associated rules of the building profile with energy consumption. This analysis will enable BCA and building owners to achieve energy efficiency that is truly effective by further optimisation.

The following parameters will be applied:

Support = 1%.

With the removal of 258 records, a total of 986 records of 1,244 are left for analysis. The minimum support value has an important role in determining the number of rules generated. As the dataset is not large, the support level should not be set too high as there will not be any rules being generated as a result. A “Support” of 1% implies an estimated 10 records to be fetched out of 1,000 records. To validate if this is a reasonable support level, a preliminary check is ran against each variable to determine that obtaining an occurrence of at least 10 for at least 1 item within the variable e.g. the item “Private School” under the variable “buildingtype” has a count of 32 (see table below for a representation). This implies that there is high likelihood that rules are obtainable.

Row Labels	Count of greenmarkyearaward
Community Hospital	4
General Hospital/ Specialist Centre (Public)	9
Hotel	308
ITE	3
Mixed Development	60
Nursing Home	35
Office	525
Polyclinic	13
Polytechnic	7
Private Clinic	9
Private College	7
Private Hospital (Private)	8
Private School	32
Retail	180
Specialist Centre (Public)	6
TCM Clinic	1
Univerisity	37
(blank)	
Grand Total	**1244**

Confidence = 30%

The dataset will be ran with a set “Support” and “Confidence” in the SPSS modeler using Apriori algorithm. Should the number of rules generated be too many, it would not be meaningful as the trivial correlations are also generated, which may cloud the analysis. As such, we will raise the support level. Conversely, if too few rules are generated, the support level should be lowered. As the Apriori algorithm is an unsupervised learning technique, it would not be possible to be able to predict the attributes that have strong correlations upfront.

An example of the mock-up model that we may expect is as follows:

Consequent	**Antecedent**	**Support %**	**Confidence %**
EUI = High	Greenmarkstatus = No and Buildingtype = Private School	20%	100%
EUI = Middle	Greenmarkstatus = Yes and greenmarkrating = Certified and Buildingtype = Mixed Development	33%	88%

The first rule suggests that many private school buildings are non-green mark building and have high EUI. This rule suggests that many mixed development which are currently holding the BCA Green Mark rating of certified (being the lowest level of certification) have middle EUI.

Evaluation

Only rules that have correlations with EUI will be used for further analysis. Broadly the two branches of analysis which could be used to evaluate the effectiveness of the BCA Green Mark scheme would be rules involving buildings that are already Green Mark certified and buildings that are non-Green Mark certified.

Deployment

The first branch of analysis related to buildings yet to apply for the BCA Green Mark scheme, would allow the policy makers of the BCA to identify who are their policies should be directed towards and how they are able to attract such building owners to apply for the BCA Green Mark scheme which ultimately achieves the outcome of low energy consumption.

The second branch of analysis will relate to buildings that are currently already on the BCA Green Mark scheme. This analysis will be able to show BCA policy makers which group of buildings (I.E Commercial/Hospitals/Schools/etc.) can be targeted to achieve higher Green Mark ratings and from there the incentives that would motivate them to upgrade their BCA Green Mark rating. This would achieve the goal of lowering the energy consumption.

In accordance with Green building incentives: A review - “In terms of effectiveness, both external and internal incentives are important instruments for promoting green building, although it is not clear which are the more effective.”, adjusting the incentive scheme is likely an effective way to promote green building.

References:

1. Green Mark Certification Scheme: https://www1.bca.gov.sg/buildsg/sustainability/green-mark-certification-scheme/ - retrieved on 16 Feb 2023
2. https://www.archtoolbox.com/energy-use-intensity/ - retrieved on 16 Feb 2023
3. Green building incentives: A review: https://www.sciencedirect.com/science/article/abs/pii/S1364032116000587 retrieved on 16 Feb 2023.