Features Affecting Airbnb Price

April 11, 2023

1 Overview

This report analysed how features affect the price of an Airbnb listing. The report cleans and implements web scraped data from Inside Airbnb, while creating new columns in progress of the research. The report determines the popularity and demand for Airbnbs and uses this justification to explore ways to predict Airbnb prices from its features. The study employs machine learning, statistical and spacial analysis to come to conclusions regarding the price of Airbnb listings in Edinburgh. The research employs RandomForest regression and OLS regression. Finally, the research looks to determine the relationship between the distance from the center of the city of Edinburgh and the price of Airbnb listings. The study confirms that there is no statistically significant relationship between the two. The study explores any extensions that coud be performed and considers the possibility of employing more data to further analyse the effect on Airbnb prices.

2 Introduction

Context and motivation Airbnb is an online platform connecting people looking to rent their houses as bed and breakfasts for people looking for short-term accommodation. The revolutionary idea has created an incredible network to boost the travel industry. It has given birth to an industry within the network; firms and individuals compete to list properties on Airbnb. However, with varying prices, features and neighbourhoods; this study will analyse how these factors are prevalent in determining a “good” property. The study will analyse these relationships by using a snapshot of the quarterly data published via Inside Airbnb.

Previous work

-

Objectives This study looks to determine what aspects of an Airbnb listing enable it to be considered a good listing. In particular, the paper will answer the following:

- How well can features of a property listing be used to predict its popularity?
- How well can features of a property listing be used to predict its short-term rental price?
- How can different neighbourhoods affect a property’s popularity?
- How can different neighbourhoods affect a property’s short-term rental price?

3 Data

Data provenance The datasets used for this project were generated by Inside Airbnb, a project founded by Murray Cox. Inside Airbnb webscrapes and consolidates public information present on the Airbnb website. The datasets are available by virtue of the Inside Airbnb data policies. The Free vs Archived Data Policy specifies that a reasonable amount of data (the last 12 months) is free for the immediate needs of the audience. A snapshot of the datasets from 16th December 2022 were made available by the University of Edinburgh. Officially the datasets are available by the Creative Commons Public Licenses – a license that gives permission to share and/or adapt the data under the terms of attribution by giving appropriate credit, providing a link to the license and indicating any changes made.

Data description This study focuses on 4 different datasets, snapshotted on the 16th of December 2022, provided by InsideAirbnb.

- **Listings Dataset** Includes 7389 rows (i.e., records) and 106 columns (i.e., features) with information on various aspects of each listing in Edinburgh. Some of the key columns are explained in the table below.

Column Explanations	
Column Name	**Explanation**
id	a unique identifier for each listing
host_id	a unique identifier for the host of the listing
neighbourhood	the name of the neighbourhood where the listing is located
lat and long	the geographic coordinates of the listing
room_type	the type of room (e.g., entire home/apartment, private room, shared room)
price	the nightly price of the listing
number_of_reviews	the number of reviews the listing has received
availability_90	the number of days in the next 3 months that the listing is available for booking
amenities	the amenities provided at a property

- **Neighbourhoods Dataset** includes 111 rows, identifying each of the neighbourhoods and their polynomial shape using geographic coordinates.
- **Calendar Dataset** has nearly 2.7 million rows depicting the availability and price of each listing in the upcoming year.
- **Reviews Dataset** has nearly 0.5 million rows for each review for a listing since January 2011.

Listings Data: To quickly and easily understand the dataset, a missingno barchart was generated. This enabled identifying the null columns (eg. *bathrooms*, *neighbourhood_group_cleansed*, *calendar_updated*), and gave a brief visual overview of the dataframe. Columns with significant proportions of missing data were removed (eg. *host_neighbourhood*, *host_about*). Checking for duplicate entries resulted in no changes.

Once the null columns and values have been eliminated, a complete dataframe of varying datatypes remained. The features in the dataframe needed to be converted into float or integer data types. Other columns that provided URLs, descriptions, sources and names that could not be converted to numeric data types were removed. Additional cleaning to remove outstanding values for certain features was undertaken. Listings with prices outside the 0.025 and 0.975 quantiles were dropped from the dataframe. Finally, of the various methods of measuring availability of Airbnb listings, the $availability\_90$ method was selected. This was due to the enforcement of regulations proposed by the Scottish Government (90 days per year rents) after significant concern from local residents and housing organisations over the increase of Airbnb and other short term lets within Edinburgh city.

Feature Engineering: Features like amenities, *host_response_time*, *amenities*, *property_type* and *room*$_t$*ype* were split into their individual components by getting their dummies and valuing each column with either a 1 (present) or 0 (not present).

Neighbourhoods Data: No cleaning was necessary.

Calendar Data: Only the *availability* and *date* columns were taken. The date column was converted to datetime format. No additional cleaning was necessary.

Review Data: Only the *date* and *id* columns were required, all others were dropped. The date column was converted to datetime format and no additional cleaning was necessary.

Once the data was cleaned, 5438 rows remained (73% of the original data).

4 Exploration and analysis

4.1 Airbnb popularity

Airbnb is a relatively recent phenomenon,launching its website in 2008 The turn of the new decade saw an immense growth, the popularity of Airbnb over time can be observed by the figure depicting the increase in the number of reviews. There is a nearly exponential increase in the number of reviews from 2012 to 2019. There is a significant drop in the number of reviews in 2020 and 2021, this is likely the result of the pandemic. However, 2022, resulted in an immense recovery with the number of reviews surpassing that of pre-lockdown numbers. This confirms the abnormality of the pandemic. The volatility of the time period for Airbnb popularity encouraged the research to focus on the price of Airbnb listings.

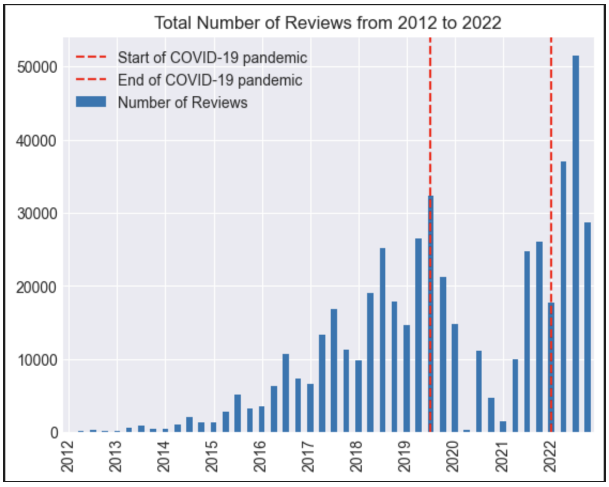


Figure 1: The total number of reviews to Airbnb Listings from 2012 to 2022

The increased number of reviews in 2022 signifies the immense increase in the demand for Airbnbs as individuals look to travel after the pandemic. This created competition in the market for Airbnb hosts after a 2-year stalemate. As Airbnb hosts look to maximise earnings in the post-pandemic world. It likely encouraged hosts to invest into their properties and provide more amenities in exchange for increased prices.

4.2 Exploring Amenities Relationship with Price

4.2.1 Hypothesis Testing

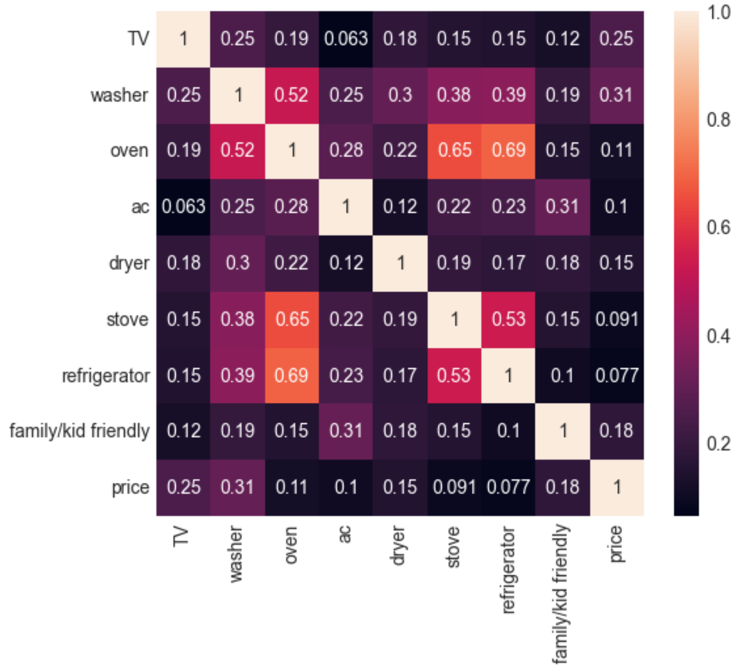


Figure 2: Heatmap of Amenities and Price of Airbnb Listing

According to this study[1], renting a Family/kid friendly Airbnb is an important concern for most families. It was shown that this feature has the highest effect on the price of the listings and according to our correlation heatmap of the amenities, we found out that tv, washer and family/kid friendly had the highest positive correlation with the price. To make sure of this, we can conduct hypothesis testing on these three amenities to check whether the presence of said amenities increases the price of a listing.

We can do this by conducting right-tailed two sample t tests [**anscombe1973graphs**] which compare the means of two independent sample groups (i.e., presence of amenity and absence of amenity) on the listing price. As t tests work best with normal distributions, a logarithmic transformation was done on price to improve the linearity between the dependent variables. This transformation would minimize the effects of outliers and provide a more accurate representation of the data. Our null hypothesis is that the presence of the three amenities will have no effect on the price of the listing and our alternative hypothesis is that the presence of the three amenities increases the price of a listing. We created a sample data frame of size n=100 to conduct these tests. We conduct this test at alpha = 0.05 with 98 degrees of freedom for each amenity as the sample size is two, making the t-critical value 1.66 (1-alpha from the t critical value table). Of the calculated t values, all of them exceeded the critical value and all the p values were below 0.05.

As the t statistics are greater than the t-critical value(t-statistic falls within the t-critical region) and as all the p values are lesser than 0.05, we have sufficient evidence to reject the null hypothesis as the presence of the three amenities increased the price of a listing.

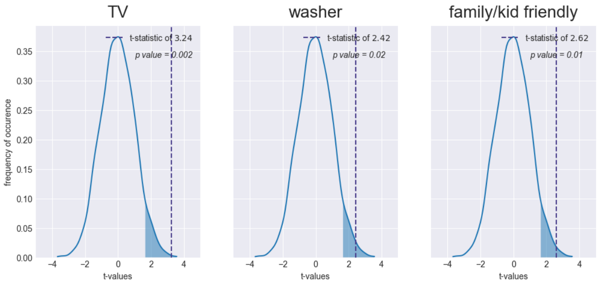
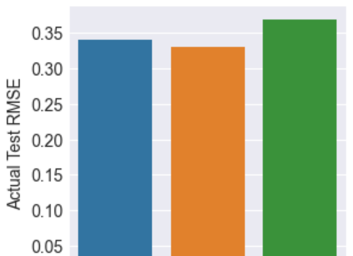


Figure 3: T-test results for amenities of Airbnb Listings with the greatest correlation with price

4.2.2 Machine Learning Models

We then wanted to figure out what all features affected the price, so we ran different machine learning models (Linear Regression, Random Forest, and K-Nearest Neighbors) . We



divided the data into training and testing data with a 70% and 30% split. For tuning the models and selecting optimum parameters, we used Grid Search cross validation which is not the most efficient as it builds a model for every combination of hyper-parameters provided but it gives a more optimized result. The coefficient of determination, or R2, is a measure that provides information about the goodness of fit of a model. In the context of regression, it is a statistical measure of how well the regression line approximates the actual data.

$$R^2 = 1 - \frac{sumsquaredregression(SSR)}{totalsumofsquares(SST)}$$

$$= 1 - \frac{\Sigma(y_i - \Theta)^2}{(y_i - \Psi)^2}$$

For the Linear Regression model, we got an r2 of 0.58, for the Random Forest model, we got an r2 of 0.6 and for the KNN model we got an r2 of 0.5. After comparing the RMSE values for each model on the testing data, we concluded that the Random Forest model secured the least RMSE value and could be used to predict the short-rental price given all the other features. Random Forest is a powerful and versatile supervised machine learning algorithm that grows and combines multiple decision trees to create a "forest." The logic behind the Random Forest model is that multiple uncorrelated models (individual decision trees) perform much better as a group than they do alone. While performing Random Forest on the data frame, we are using the Gini index, or the formula used to decide how nodes on a decision tree branch.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

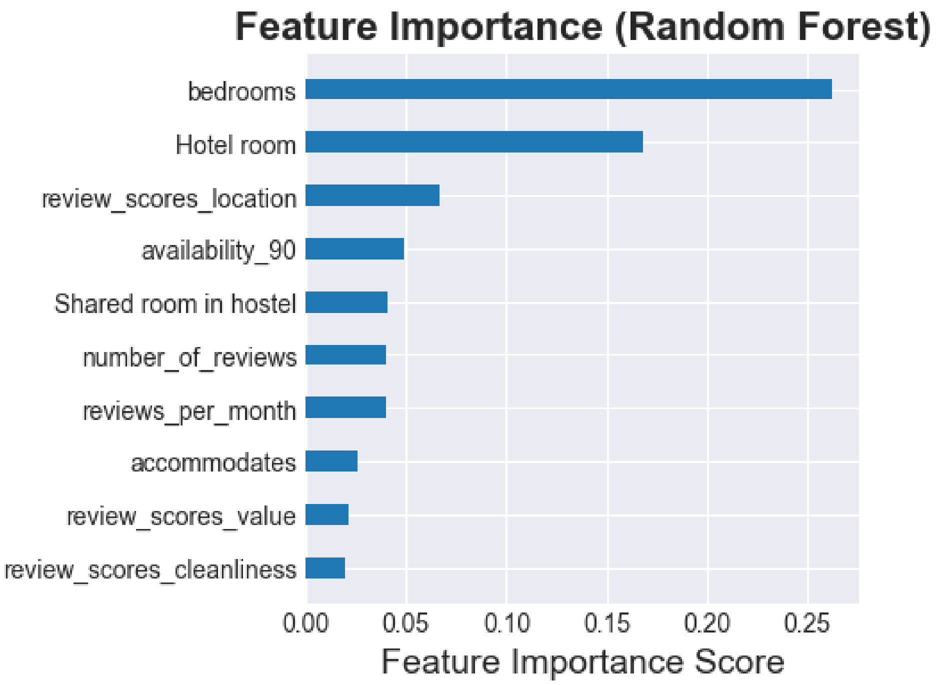


Figure 5: Most important features in predicting price of Airbnb

The Random Forest algorithm has built-in feature importance which can be computed using Gini importance. For each feature we can collect how on average it decreases the impurity and the average over all trees in the forest is the measure of the feature importance. We have applied this feature to visualize the most important features according to the random forest model. According to the model, number of bedrooms is the most important feature followed by Hotel room and then by the review scores by location.

4.3 Neighbourhood Discoveries

The study also conducted spacial analysis by plotting the listings data on two maps of Edinburgh.

Figure 6 depicts every individual listing in Edinburgh, every listing is represented by a dot and is coloured in relation to the price of the listing. It is evident that there are a lot of listings in the center of the city, while decreasing in density as one moves towards the countryside. This coincides with the understanding that the city center offers more purpose and attraction to tourists. It is interesting to note the role that some listings play on influencing the average price of the neighbourhood.

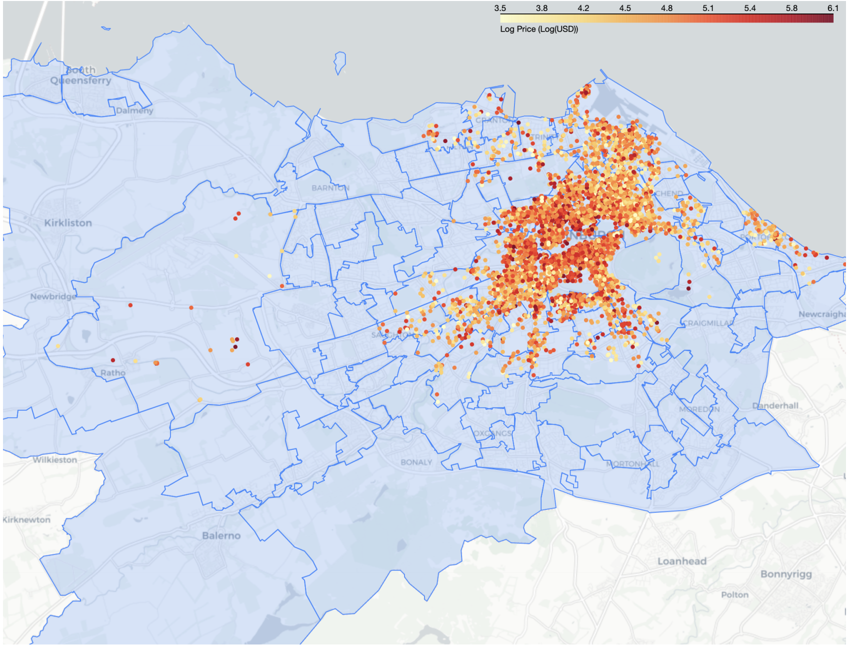


Figure 6: Listing of each

The exception can be seen through the space of emptiness around Arthur's seat and Hollyrood Park, however, the neighbourhood around still results in a significant number of listings as seen in Figure 7. This signifies the importance of the central location and the demand for visitors to be near the heart of the city. It is clear that the price ranges from as low as $42 and peaks at $168. However, a more interesting phenomenon is the diversity of the prices within a neighbourhood as well as between neighbourhoods entirely. There is visible variation in the prices of Airbnb listings around the city, besides the center.

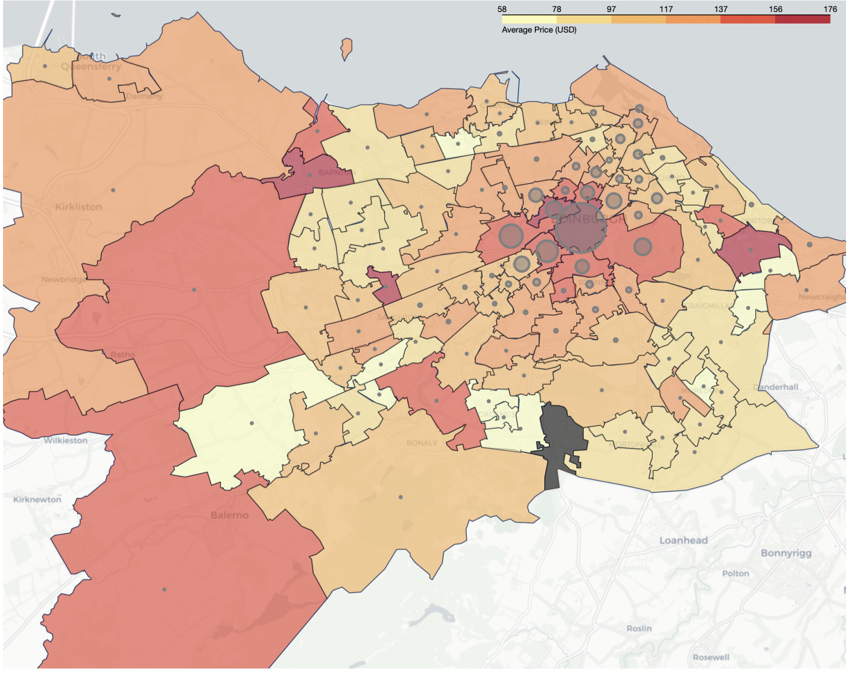


Figure 7: Heatmap of Amenities and Price of Airbnb Listing

Comparing the number of listings to the neighbourhood average price puts the relationship into perspective. There are so few listings in some neighbourhoods like Bonaly, Balerno and Kirkliston that the data could be skewed. This is evident in cases like Barnton, Cammo and Cramond South; Carrick Knowe and Mountcastle who have fewer than 10 listing each, yet have high median listing prices. This creates additional doubt regarding the usability of the data from the outskirts.

To the eye, it seems obvious that the flats near the city center are priced higher. However, to confirm if the location of an Airbnb hosting in Edinburgh has a significant effect on the price of an Airbnb listing, the data needs to be tested.

4.4 Effect of Distance on Price

To examine the effect of the distance on the price, a sample of data that eliminated the outliers. Listings from neighbourhoods with fewer then 20 listings were removed as they would cause inaccuracies in the testing and regression. There needed to be a focal point from which the distance of each listing could be calculated. A bollard in the middle of Edinburgh, identified and coined by Atlas Obscura as 'the center of Edinburgh' bollard was an appropriate location to consider the center. It stands opposite Waverly station at the coordinates (55.9531, -3.1882) [**https://www.atlasobscura.com/places/center-of-edinburgh-bollard**].

- H_0: The null hypothesis states that the distance of an Airbnb listing from the identified city center (center bollard) does not affect the logarithmic price of an Airbnb listing.
- H_1: The alternative hypothesis states that the closer distance from the identified city center (center bollard) causes an increase in the logarithmic price of an Airbnb listing.

The distance from the bollard to the individual Airbnb listings, is calculated using the radius of the earth at 6371 kilometers and performing trignometric algebra on the two latitude and longitude coordinates[2]. This enables the creation of a new column for the distance from the bollard.

This relationship is examined using OLS regression. Unfortunately, the results of the regression indicated that the model explains little to none of the variability of the price against the distance to the bollard. Although the p-value indicates absolute significance with a p-value of 0, the R^2-value was too small to prove any significant relationship.

Dep. Variable:	y	**R-squared:**	0.037
Model:	OLS	**Adj. R-squared:**	0.037
Method:	Least Squares	**F-statistic:**	195.1
Date:	Tue, 11 Apr 2023	**Prob (F-statistic):**	1.56e-43
Time:	10:19:35	**Log-Likelihood:**	-3794.9
No. Observations:	5090	**AIC:**	7594.
Df Residuals:	5088	**BIC:**	7607.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	4.8900	0.013	382.785	0.000	4.865	4.915
x1	-0.0825	0.006	-13.967	0.000	-0.094	-0.071

Omnibus:	15.672	**Durbin-Watson:**	1.855
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	13.453
Skew:	0.066	**Prob(JB):**	0.00120
Kurtosis:	2.786	**Cond. No.**	4.47

The regression was then conducted with the RandomForest Model; however, this resulted in a p-value of 0.7, which indicates that it is statistically insignificant. The regression was also conducted with the DecisionTreeRegressor Model; however, this resulted in a p-value of 0.99, which was also statistically insignificant.

This forced us to come to the conclusion that the distance from the city center does not statistically affect the price of an Airbnb listing. Therefore, the null hypothesis cannot be rejected.

5 Discussion and conclusions

Summary of findings Using hypothesis testing, we were able to see that the presence of certain amenities (TV, washer, family/kid friendly) increased the price of a listing.

On creating the various machine learning models, we found out that the random forest model predicted the price the best with the number of bedrooms as the most important feature.

Evaluation of own work: strengths and limitations Strengths: The report incorporates a variety of machine learning techniques and employs hypothesis testing. It implements methodology observed and practiced regularly in the data science community. The report includes appropriate and relevant background information regarding the models and methods used. There are a variety of visible and comprehensive visualisations used to effectively communicate the results and findings.

Limitations: The report could have produced more progressive results with regards to the outcome of the report. This is most significantly the case with the lack of significance in the relationship between the distance from the Airbnb listings and the center of the city of Edinburgh.

Comparison with any other related work We found to have an opposite conclusion to the paper about predicting price [3]. The authors of the paper concluded that the number of accommodates was the most important feature in the models but according to our results, we found that the number of bedrooms was the most important feature.

Improvements and extensions There are many imrovements that could ahve been made to the report. Majorly, another dataset could have been taken alongside the InsideAirbnb dataset. This could have provided information about the weather, or large events in Edinburgh over the year, or other information. This would have made the report more interesting to examine and could have resulted in more effective and concrete findings. Before doing any regression, the report could have conducted hypothesis testing on the coefficients of the report. A more comprehensive model could have been generated, that incorporates all the available data ad bootstraps the data according to different superclasses of features. We believe these improvements and extensions could have been made given more time.

References

[1] Rahul Gupta. *Identifying most important amenities affecting the price of Airbnb*. Sept. 2019. URL: https://medium.com/analytics-vidhya/identifying-most-important-amenities-affecting-the-price-of-airbnb-3a34af7e4a38.

[2] Yifei Jiang et al. "How to better incorporate geographic variation in Airbnb price modeling?" In: *Tourism Economics* (2022), p. 13548166221097585.

[3] Siqi Yang. "Learning-based airbnb price prediction model". In: *2021 2nd International Conference on E-Commerce and Internet Technology (ECIT)*. IEEE. 2021, pp. 283–288.