International Journal of System Dynamics Applications
Volume 10 • Issue 4 • October-December 2021

Property Valuation Using Linear Regression and Random Forest Algorithm

Sam Goundar, The University of South Pacific, Suva, Fiji
https://orcid.org/0000-0001-6465-1097

Akashdeep Bhardwaj, University of Petroleum and Energy Studies, Dehradun, India
https://orcid.org/0000-0001-7361-0465

ABSTRACT

The economic boom over the recent past and the quest to further develop has made several nation states the business hubs in their regions. Along with the investments, there has been growth in the number of property sales. Social media has become convenient platform of choice for advertising property sales after the introduction of Web 2.0. This article utilizes social media platforms like Facebook to scrape data from user groups advertising properties and then using data mining techniques and approaches to determine true valuation of properties. This methodology is based on set attributes in the urban areas by looking at the property sales of the recent past within the same area. This enables investors interested in these properties and provides a fair idea of price of properties based on the key attributes associated with the respective property.

KEYWORDS

Data Mining, Linear Regression, Property Valuation, Random, Social Networks

1. INTRODUCTION

Data mining provides information from data sets that could be utilized by drawing comparisons, associating patterns, classifying objects, to predicting future trends, based on the current and past standards of any given situation provided a clear set of data. Realizing the importance and the scope of mining for information, this article provides focus and predicts the market values of properties situated in major urban locations on mainland Fiji that are being advertised on the social media platform, Facebook, based on the current and the past sale of properties within these respective areas. The introduction and progression towards Web2.0 have transformed the use of internet, enabling buyers with the ability to comment, review, blog, update status amidst many other forms of interaction and active participation. One of the new revelations (Branko et al., 2013) of such usage and participation has been the introduction of big data, with diverse, unstructured stream of information floating on the web containing blends of data such as demographics, lifestyle choices, opinions, as well as mentions of properties and items in possession by the people, among many others. The authors choose Facebook as the social media platform to scrape data from, as it is the most common social media platform of

DOI: 10.4018/IJSDA.20211001.oa13

choice in Fiji, and also since Facebook provides user groups based on common interests (in our case, property sales groups created to advertise properties).

Too often in real estate, (Geojournal, 2020) the process of valuation can come across as a high-brow exercise of thumb-sucking. The realtor will come over, kick the proverbial tires, and then produce an estimated value with very little "quantitative" insight. Perhaps the process is exacerbated by the emotional attachment that owning property brings given that for many, a house will be the largest financial investment made in a lifetime. The comparable property valuation approach is most common model for determining residential real estate and recent sales of similar or properties to determine the valuation of a subject property. The sales price are adjusted based on differences between these and the subject property. For example, if a comparable property has an additional bathroom, then the estimated value of the bathroom is subtracted from its observed sales price. Real estate is considered to be more heterogeneous, so the comparable valuation approach is used less frequently (Doumpos et al. 2020). The income approach, based on the concept that the intrinsic value of an asset is equivalent to the sum of all its discounted cash flows, is more commonly applied across two methods. The first is similar to the present value of an annuity, the direct capitalization method uses the net operating income (NOI) of a property divided by the "cap rate" to establish a value. The cap rate contains an implied discount rate and future growth rate of net operating income. While the second method involves discounted cash flow method provides the present value of future cash flows over a set period of time, with a terminal value that is estimated from using a terminal cap rate. The final technique is the cost approach, which estimates value based on the cost of acquiring an identical piece of land and building a replica of the subject property as presented by José-Luis et al. (2020). Then cost of the project is depreciated based on the current state of obsolescence of the subject property. Similar to the adjustments in the comparable sales approach, the goal is to closely match the subject property. The cost approach is less frequently used than the other two approaches.

The aim of the to identify texts from structured as well as unstructured posts from social media, which infer to properties such as lands and houses attached with variables such as price, along with mentions of other details and attributes. The research intends to further the work and evaluate by annotating these texts and analyzing the identification, and storing these identifications with its variable in an updated dataset. Use of data mining techniques such as linear regression, and random forest model is performed to determine the predicted valuation figure and testing it with the current valuation reports on the selected properties which can be used to compare against prices advertised on the social media groups.

The motivation for this research is led through the fast progression and investment opportunities opened due to the economic boom that the country is currently enjoying in its quest to further develop towards being the business hub of the South Pacific, due to factors such as political stability and overseas investors. One of the main outcomes of these are investments in the country, and among the investments is the growth in the number of property sales, greatly involving real state agencies and property values. Valuation is a common measure taken by the investors to protect and ensure the property being invested into is within the range of the market rates. However, with users having the ability to post and share direct information of properties on social media platforms, people looking to invest are drawn to these properties without any fair knowledge of its value, often opting not to involve consultants, licensed agencies and professional values to determine the current valuation of the properties they seek. This article begins with an overview of related literature, followed by the methodology adopted in scraping and annotating text from users' posts on Facebook user groups specializing in adverts on property sales, where we highlight the adopted methodology utilized in determining properties on sale in major towns and cities on mainland Fiji. The variables and attributes extracted from each of the posts and the considerations needed to ground the property annotation process, ultimately result in a proposed annotation schema.

The authors introduce the data scraped from the social media platforms' user groups and the various attributes pertaining to the dataset. The research focused on modelling the data by pre-processing and mining to determine the variables as factors that affect sale prices on the properties. By using real estate listings with valuation reports as a supervised learning problem, from which the prices advertised are compared by users on Facebook against the known listings as unsupervised problem to predict the valuation on the properties being advertised. The users also seek relations between the prices being advertised and the possible valuation, along with pertaining errors in our machine learning case. Finally, the authors present the conclusions of the findings as summary, whilst highlighting probable future work henceforth.

2. LITERATURE REVIEW

Cultural productions are considered as a sign of civilization in modern societies. Theater is known as an important type of cultural productions, playing important role in the cultural economy of a society. Due to complexities of socio-economic interactions, this sector needs dynamic investigation to illuminate different aspects of possible potentials and threats. Abdollahi et al. (2019) presented relationships between Iran public theater economy and production structure based on a dynamic model including all economic stages, namely production, distribution, and consumption to achieve a solid perception of Iran theater position. The authors use System Dynamics to create a model that can explain or mimic the behavior of the system in order to evaluate policies. Since Tehran City Theater complex is the sole place for the public theater in Iran, the authors assess it over the period 2012-2015 and predict its behavior to 2022. On the other hand, the investigation in this context is being directed in accordance with microeconomics principles. The results indicate that the position of Iran public theater is undesired due to vague managerial policy. Also, the findings offer insights into the problems and suggest practical solutions.

Task administration is the procedure through which work is instated, arranged, executed, and controlled by a group to accomplish an objective. Since project management activities are different from normal business activities that are conducted every day, project management calls for special technical and management skills amongst team members. The successful completion of a project depends largely on systems engineering and the management of various programs. Systems engineering refers to an interdisciplinary approach that facilitates the realization and success of complex systems. Galli, B. (2019) proposed application of system engineering to project management. The purpose of system engineering is to, therefore, influence the whole system through various cohesive subsystems. The principles of system engineering are synonymous with most characteristics of project management. System engineering is, therefore, applicable to advance project management.

Shukla et al. (2019) focuses on the environmental, economic and social impact of marble industries in the north-western region of India. The research presents a grey-based decision-making model for evaluating the extent of sustainability in three marble processing industries. The goal of this article is twofold. First, to identify the important criteria of sustainable performance in marble sector and second to compare three marble processing firms on the basis of sustainability criteria using grey based decision-making approach. A detailed questionnaire was sent to three marble processing firms and the analysis is done on the basis of the received responses.

Software is an important part of human life and with the rapid development of software engineering the demands for software to be reliable with low defects is increasingly pressing. The building of a software defect prediction model is proposed in this article by using various software metrics with publicly available historical software defect datasets collected from several projects. Such a prediction model can enable the software engineers to take proactive actions in enhancing software quality from the early stages of the software development cycle. Panda, M. (2019) introduced a hybrid classification method (DBBRBF) by combining distribution base balance (DBB) based instance selection and radial basis function (RBF) neural network classifier to obtain the best prediction compared to the existing research. The experimental results with post-hoc statistical significance tests shows the effectiveness of the proposed approach.

Herrera et al. (2019) proposed customer value generation has drivers, which could be different according to each stakeholder within the electricity industry, affecting its growth. Each stakeholder has different interests that affect the decision-making process and the customer value perception in the long term, which impacts on profitability. In order to illustrate how to identify and model key performance drivers to evaluate creating value in the electricity utility industry, this study used a simulation with the system dynamics methodology. Through simulation scenarios, this study shows that, the high customer value perception allows the electricity utilities industry to create more value. This is illustrated with the case of some electricity utilities engaged in the generation and distribution in the Colombian electricity market. The results show a new point of view that contributes to marketers and engineers in the analysis of the relationship between the stakeholders and electricity firms.

Ghabi et al. (2018) introduced a sliding mode controller to stabilize a discrete-time nonlinear system in the presence of uncertainties and external disturbances. The proposed controller is derived to guarantee the existence of a quasi-sliding mode, taking into account the upper bound of uncertainties. With this method, a recursive switching function is used, which allows for recovering lost invariance and robustness properties of a discrete sliding mode control. As for the system stability, it is found that the system is stabilized and finally restricted to a known region. This control scheme ensures robustness against parametric uncertainties and external disturbances as well as the elimination of chattering. In this article, after a detailed formalization of the proposed control design, a numerical example for an inverted pendulum is considered, proving the effectiveness of the control methodology.

Vibration control of fractional-order linear systems in the presence of time delays has been dealt in this article. Considering a delayed n-degree-of freedom linear structure that is modeled by fractional order equations, a fractional-order optimal control is provided to minimize both control input and output of delayed system via quadratic objective function. Balochian et al. (2018) first the fractional order model of system that is subject to time delay is rewritten into a non-delay form through a particular transformation. Then, a fractional order optimal controller is provided using the classical optimal control theory to find an optimal input control. A delayed viscose system is then presented as a practical worked-out example. Numerical simulation results are given to confirm the efficiency of the proposed control method.

Chaotic behavior is a term that is attributed to dynamical systems whose solutions are highly sensitive to initial conditions. This means that small perturbations in the initial conditions can lead to completely different trajectories in the solution space. These types of chaotic dynamical systems arise in various natural or artificial systems in biology, meteorology, economics, electrical circuits, engineering, computer science and more. Of these innumerable chaotic systems, perhaps the most interesting are those that exhibit attracting behavior. Moysis et al. (2017) referred to systems whose trajectories converge with time to a set of values, called an attractor. This can be a single point, a curve or a manifold. The attractor is called strange if it is a set with fractal structure. Such systems can be both continuous and discrete. This paper reports on some new chaotic discrete time two dimensional maps that are derived from simple modifications to the well-known Hénon, Lozi, Sine-sine and Tinkerbell maps. Numerical simulations are carried out for different parameter values and initial conditions and it is shown that the mappings either diverge to infinity or converge to attractors of many different shapes.

Hussein et al. (2017) presented a method for fault detection and diagnosis of stator inter-turn short circuit in three phase induction machines. The technique is based on modelling the motor in the dq frame for both health and fault cases to facilitate recognition of motor current. Using an Adaptive Neuro-Fuzzy Inference System (ANFIS) to provide an efficient fault diagnosis tool. An artificial intelligence network determines the fault severity values using the stator current history. The performance of the developed fault analysis method is investigated using Mat lab / Simulink® software. Stator turns faults are detected through current monitoring of a 2-HP three phase induction motor under various loading conditions. Fault history is calculated under various loading conditions, and a wide range of fault severity.

Big Data analytics involves various technique to exact data, analyze and predict patterns in data. This concept helps organizations to extract quality information from data for decision making. It is considered as a collection of tools and techniques that work together to uncover the unseen relationships between data (Smith et al. 2011). To be more precise Data Mining is a field of computer science that helps in discovering patterns and extracting hidden information to form organized formats of useful information. Data mining involves the use of artificial intelligence, machine learning, statistics, and databases (Leventhal, 2010).

To date, research focusing on extracting latent user-descriptive attributes from microblogs has been mostly centered around Twitter, as it is a service with high adoption rate, where many of the users share tweets publicly (Banea et al., 2018) on attributes targeted for extraction such as demographics, gender and age. Scraping for data from web, especially on social media platforms, can be very useful as it contains a wealth of information in structured and unstructured form. However, the scraping process can become very challenging as most of these platforms, in order to maintain user privacy and confidentiality, secure and restrict web scraping. As in the case of this research article, the needful web scraping for data on property listings was reduced to manually searching for related posts, as Facebook, since 2016 has restricted all forms of data scraping enabled to users under its developer platform.

A major impediment in conducting research in this area has been the lack of training data (Cheng et al. 2010) since information in user profiles such as location, gender, age and ethnicity are leveraged from being self-specified. As reported by Banea et al. (2018), only 26% of the users on social media specify their true information or information which has not been altered to any extent. As such, human annotation effort becomes necessary to validate data obtained on the information from online social platforms. As a complementary method, Volkova et al. (2016) have employed predictions made by machine learning algorithms trained on a smaller set of data enhanced with socio-demographic user attributes identified through crowd sourcing annotation task, to generate a larger training set, which after subsequent trainings, yield results more precise than single stage learning algorithm (Banea et al. 2018). Such an approach to information pertaining on online social media platforms yields state of an art results in terms of information, despite the lack of predicted accuracy.

To date, several cases and project of similar nature has been embarked upon and are built around from many parts of the world, such as Spain, Chile, and Australia, where different data mining techniques have been utilized in hybrid form on data obtained from real estate agencies. According to Irfan et al. (2015), social networking websites are very rich in texts that enable user to create various text contents in the form of tweets, wall posts, comments and so on. The most popular social networking sites currently are Twitter, Facebook, LinkedIn and MySpace containing information in structured, semi-structured and unstructured forms. Text mining can be applied to extract data from networking sites and provide a knowledge discovery process by forming multidisciplinary fields as attributes. Most of the scientific literature, Xu et al. (2008) focuses on specific techniques on text mining for information extraction, however, a thorough discussion is lacking on actual analysis of different text mining approaches.

Valuation plays an important role and is crucial in deducing current property value, which safeguards the customers and potential buyers from being victimized to prize hikes. Trojanek, (2010) states the many implications poised at lack of valuation and underlines the basic procedures involved to determine the factors involved in deducing current values. Forecasting real estate prices can be a challenge and different methods prevail, each with its own set of obstacles. Text mining algorithms and linear regression model was utilized by Khashan, (2014) in the quest to forecast real estate prices in Dubai. The author managed to arrive to the desired results with minimal errors. We will be utilizing a similar approach in the initial task, heavily relying on linear regression and moving on to predict valuation using random forest algorithms.

Random forest is a clustering technique built around complex decision trees. Ali et al. (2012) recommend the random forest classification and decision tree algorithm in extraction, classification and comparing data. In their approach, however, they also encounter setbacks with sorting of numerical attributes. Murrel, 2014 elaborates that decision trees take a greedy approach and may not always be able to provide the best decision tree. The entropy utilized in itself may provide a bias result, hence, more often than not, other techniques such as artificial neural networks are preferred over the random forest classification techniques. A rigorous study on artificial neural network delineated the use of time series data and incorporation of past trends; successful implementations of which enable forecasting the chosen variable. As far as this research problem is concerned, the best suited model for making predictions based on current prices is random forest clustering.

In addition to text mining, regression analysis can cause prediction errors to be identified and minimized. Decision tree and Regression Splines methods as mentioned by Acciani et al. (2011) accomplished better results even with small datasets. To some extend small datasets do not reveal an accurate estimation in terms of predicting the accuracy, the researchers found that Model tree and Regression Splines needs further research. Best practices and integrated models require numerical data, hence, Saravanan et al. (2010) state that where possible, the character data be transformed in logical numerical values. Incorporation of random forest with logical regression enable smooth investigation unstructured data. A huge number of pre-processing techniques can be applied to text mining in relation the need of the survey or the type of related data needed for prediction. In addition to, a prior step of preparing text content for text mining process should be taken into account. Several pre-processing methods can be applied according to the needs and the purpose of text mining. Clustering methods utilizing decision trees such as Random Forest algorithms has been utilized by Cacho et al. (2010) produce effective and accurate results. Jaen, (2002) built an algorithm, named CARANS, for analysis and prediction for quantity evaluation in urban areas which produced a 15% abnormality to the quoted price of houses.

In doing this project, several articles and online research was utilized, which motivate the use of the techniques mentioned utilizing data mining for prediction and determining variable factors in big data. Several successful research and literature support that property valuation is a concern where intelligent use of models can be utilized to avoid unnecessary fluctuations in property prices. Data mining is an essential tool to cater for variations in property pricing.

3. RESEARCH METHODOLOGY

The authors separate the objective of determining valuation of properties obtained from the user groups on Facebook, into two tasks. In order to test the dataset constructed from posts, different dataset is created which comprises valuation on properties obtained from real estate agencies' listings from around the country. This dataset was used to train the proposed model using linear regression techniques of data mining. The sample of this train dataset comprised of 420 records with variables as enlisted below.

In the quest to maintain accuracy with valuation and recency, the authors limited the dataset with listings of the past six months. The relatively small sample size can also be attributed to the overall population on mainland Fiji. Having trained the model with this dataset, a separate dataset, created from Facebook groups was used to test and predict the valuation prices and the predicted valuation compared with the prices of properties advertised to deduce unknown relationships. Table 1 shows the variables used to build the dataset.

3.1 Feature Extraction and Description of Attributes for Variables

- **Land Tenure:** Despite Fiji being a republican nation, the land tenureship does not belong to the state but is divided as native and freehold land. For this research, the focus is on residential properties, the authors emphasize only two attributes towards the Land Tenure variable, '0' denoting native land and '1' denoting freehold land. Other types of land such as agricultural

Table 1. Variables to train our model to determine valuation of properties

Variables	
1	Land Tenure
2	Location
3	Price
4	No. of Bedrooms
5	Lease Title
6	Engineers Certificate
7	Lot Size
8	Garage
9	Air Conditioning Units
10	Swimming Pool
11	Flats
12	Washrooms

does not apply to this study as the research focuses solely on valuation of properties in sub-urban areas. Also, industrial and commercial type is eliminated as the focus on residential lots.

- **Location:** The properties extracted for this study are all situated in urban areas around mainland of Fiji. The geographical demographics of this nation is based on the coastline around the main island, hance our focus has been around major towns and cities, namely, Lautoka, Nadi, Suva, and Nausori corridors.
- **Lease Title, Engineers Certificate, Garage, Air Conditioning Units, Swimming Pools, Bathrooms:** These attributes were converted to binary for better interpretation and modelling using artificial neural network. ‘0’ denoted unavailability and a ‘1’ stood for availability of these respective amenities and documents.

Having attained the dataset for training the model, the authors perform pre-processing, initially by eliminating variables other than the ones mentioned above. The authors changed the values of the data into logical data for variables where the attributes require numerical data. As far as the missing values are concerned, k-means clustering is selected to deduce nearest neighbour, hence, obtaining values for the attributes. K-Fold cross validation technique has been implemented in preparation of the training set for validation. The authors divide the training set into five folds, as each of the entry in the validation set is used once. Taking k as 5, the five folds deduce different cross-fold examination result, and based on the outliers, the authors determine the best of the five validated results to be utilized for training the data.

To train the data with minimal errors, Root Mean Square Error value is used of the initial model and compare the error with a second model, having eliminated from the train dataset, the variable with negligible impact on the valuation price. Once the model with lowest error value is determined, the results will be applied to the test data to predict valuation on property prices obtained from the scraped data on social media and a comparison is drawn with the previous available properties sold and valuation records to determine the validity of the predicted valuation.

The authors include the test data, where the valuation value is set to void and from the trained dataset, deduce the best value for the valuation, depicted along with error analysis in the following results section. The data obtained for this second task was obtained from user groups on Facebook which are specifically created for property sales in the locations of our interest. 13 different such groups are utilized and managed to obtain 123 adverts that form records in the dataset. Extraction of data from these user groups had to be done manually, as the extraction and text mining techniques mentioned in the literature earlier proved futile due to the social media security perimeter and policies in place. Each of the records extracted from these groups were manually annotated and updated in the dataset. As expected, not every listing advertised shared all the attributes that formed the variables, hence pre-processing and handling of missing data is performed for nearly each of the records attained. The train dataset focused on data obtained from real estate agencies, which were dating back to last 6 months, hence, in order to keep the predicted valuation accurate, this research is limited to the extraction of data from social media to the past 6 months as well. It also ensured that the error margin is kept to a minimum.

Having prepared the dataset, the authors perform a hybrid random forest modelling of the data. Identified variables are utilized which impact valuation prices obtained from the linear regression model developed and perform pre-processing on the data developed from social media posts by eliminating variables which do not impact the valuation and property prices. Having done so, the model is trained using the real state data and predict property valuations on social media data. The observations and predictions made are discussed in the next section.

The reason for choosing regression analysis is foremost it handles multiple things, e.g. use of regression analysis can perform multiple independent variables for models, utilize polynomial terms to model curvature, include continuous as well as categorical variables and even assess interaction terms. This can help determine the impact of independent variables that are depending on the value of another variables. Such abilities of regression analysis also include the magical ability of unscrambling intricate problems, in real-world use cases when multiple variables are combined.

4. RESEARCH PERFORMED

As discussed earlier, the research objectives set out to achieve two critical tasks in the quest to predict and enable potential and interested buyers and general public access to true valuations. As per the first task, it becomes important to deduce factors which affect the price of properties from a whole range of variables which get advertised in marketing of an individual property, such as the land type, the type of property, whether the property has documentations such as proper title and engineers certificate, if applicable, and factors such as bedrooms, garage, swimming pools and so on which easily lure buyers into potential reasoning. The authors set out to determine which amongst these many factors truly affect the valuation and pricing of properties in Fiji. Figure 1 below shows the number of properties and the valuation price based on the listings obtained from real estate agencies. Most of the records in the listings are under $FJD 100,000.

Several other factors such as price distributions specific to locality were studied to better enable the designing of a proper linear regression model. Figure 2 below shows the price distributions along the major urban locations on the main island. From the figure, the authors are able to deduce that most of the data is centralized amongst a quarter million dollars with few properties hitting over half million dollars.

The authors are also able to visualize several other important information such as the inter-dependency of several variables in deducing the property prices. As an example, Figure 3 below uses the number of bedrooms and the area size as two inter-dependent variables and how prices are affected by these variables.

5. RESULTS

Having visualized the dataset, the data is split into training and test sets with a split ratio of 0.7. Out of the 440 recent properties, dating back to the past six months, the authors build out training data by utilizing 308 records and the other 132 records as the test data. These models effectively generate

Figure 1. Properties and Valuation based on Real Estate Listings

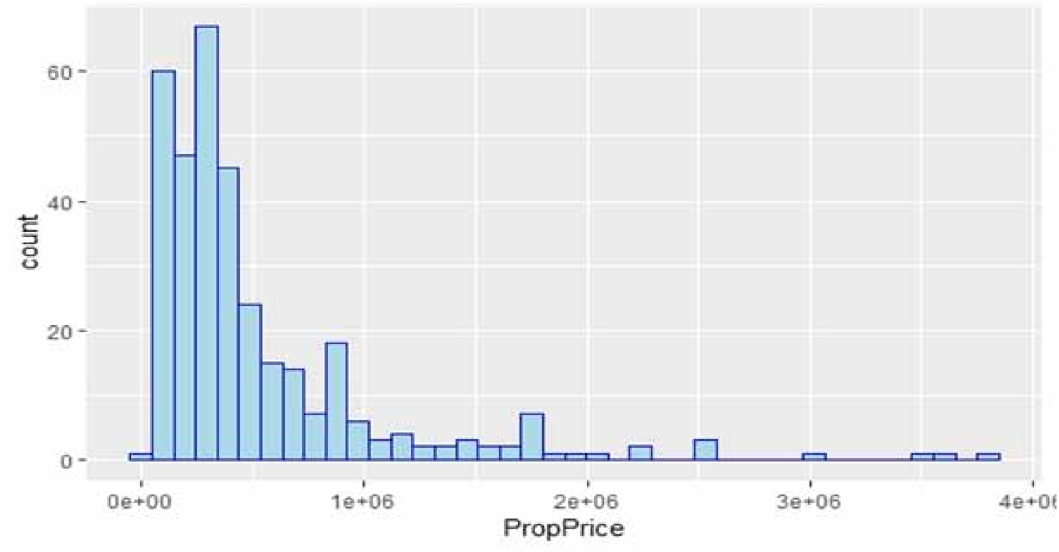


Figure 2. Price range specific to Major Urban Locations

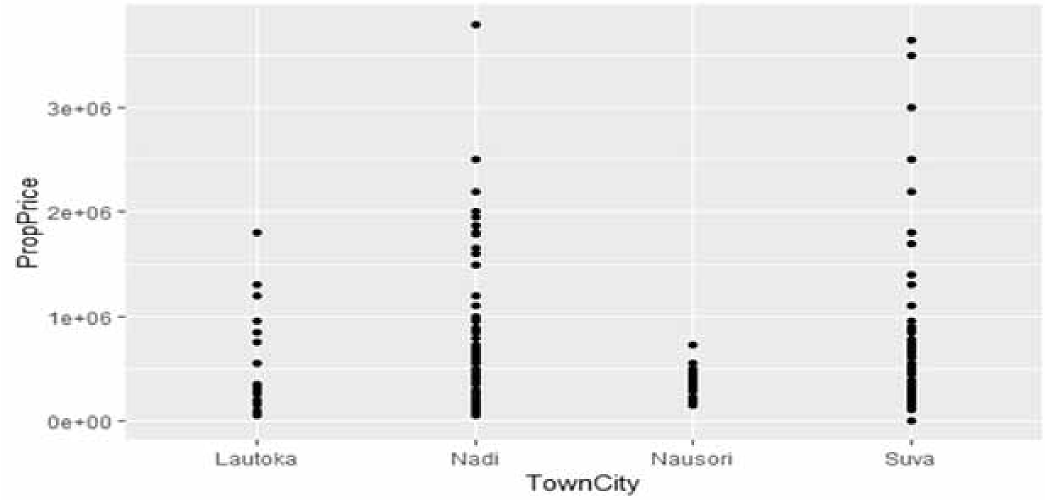
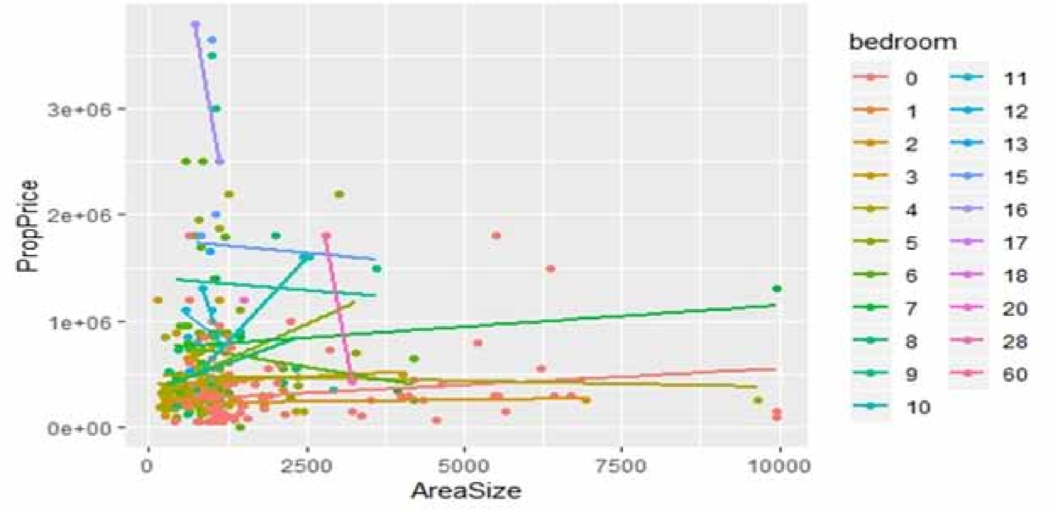


Figure 3. Inter-dependencies between Variables



the linear regression model. The following result obtained highlights variables, indicated by asterisks (the greater the number of asterisks, the higher the impact of the respective variable on the valuation price) which impact the valuation prices in Table 2.

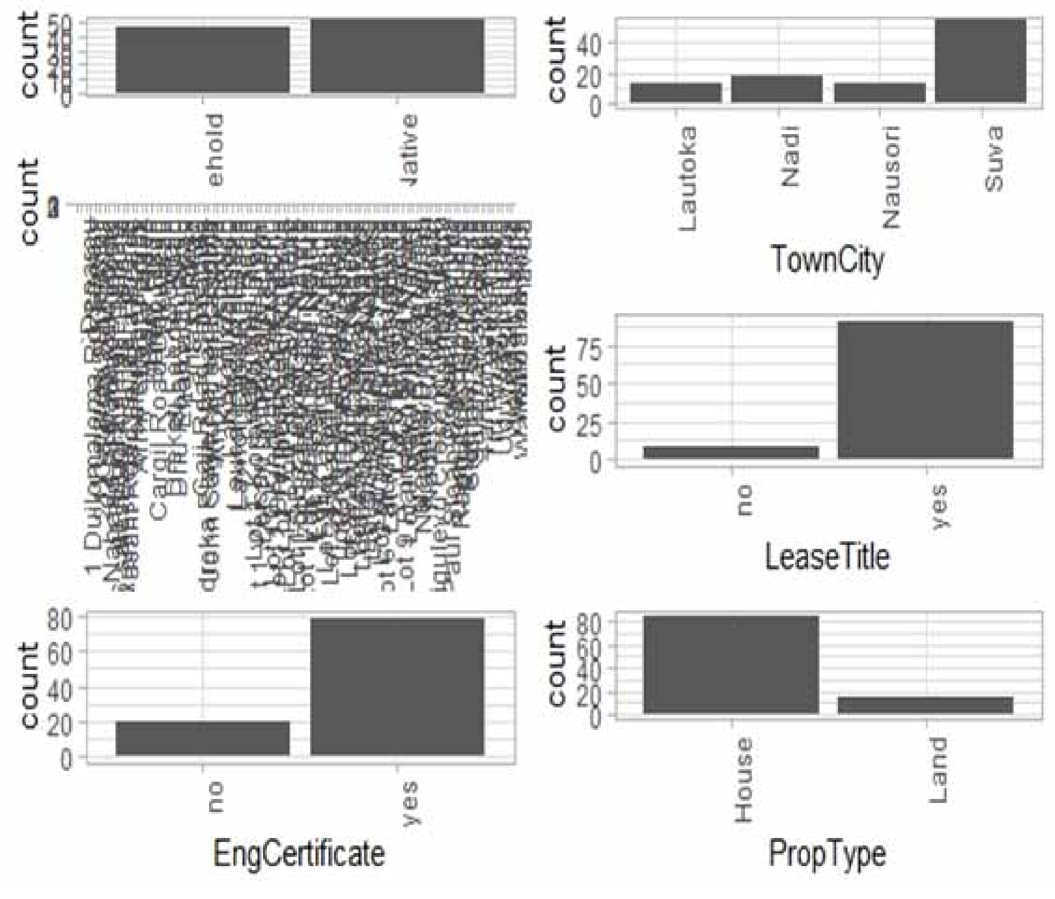
Table 2. Variables and Level of Impact on Price Valuations

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.197e+05	2.216e+05	0.991	0.32250
Description	-1.401e+05	6.986e+04	-2.005	0.04611 *
TownCity	2.904e+05	1.448e+05	2.007	0.04597 *
LeaseTitle	3.077e+04	9.359e+04	0.329	0.74265 **
EngCertificate	8.239e+04	1.090e+05	0.756	0.45037 *
PropTypeLand	4.437e+04	1.441e+05	0.308	0.75835 *
AreaSize	9.042e+00	2.217e+01	0.408	0.68372 **
NumFlats	-6.971e+04	2.844e+04	-2.451	0.01498 *
Bedroom	9.938e+04	1.862e+04	5.337	2.26e-07 ***
Garageyes	-1.325e+05	1.011e+05	-1.311	0.19129
SwimmingPool	4.417e+05	1.337e+05	3.302	0.00111 **
FullyFenced	-1.032e+05	9.890e+04	-1.043	0.29802
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’ 0.1 ‘ ’ 1

```
Lm (formula = PropPrice ~ data = train)
Residuals:
   Min        1Q      Median      3Q        Max
```

```
-1320878  -206363   -96655    75927   2531574
Coefficients:
Residual standard error: 506600 on 231 degrees of freedom
Multiple R-squared:  0.3322,
Adjusted R-squared:  0.2946
F-statistic: 8.838 on 13 and 231 DF, p-value: 1.289e-14
```

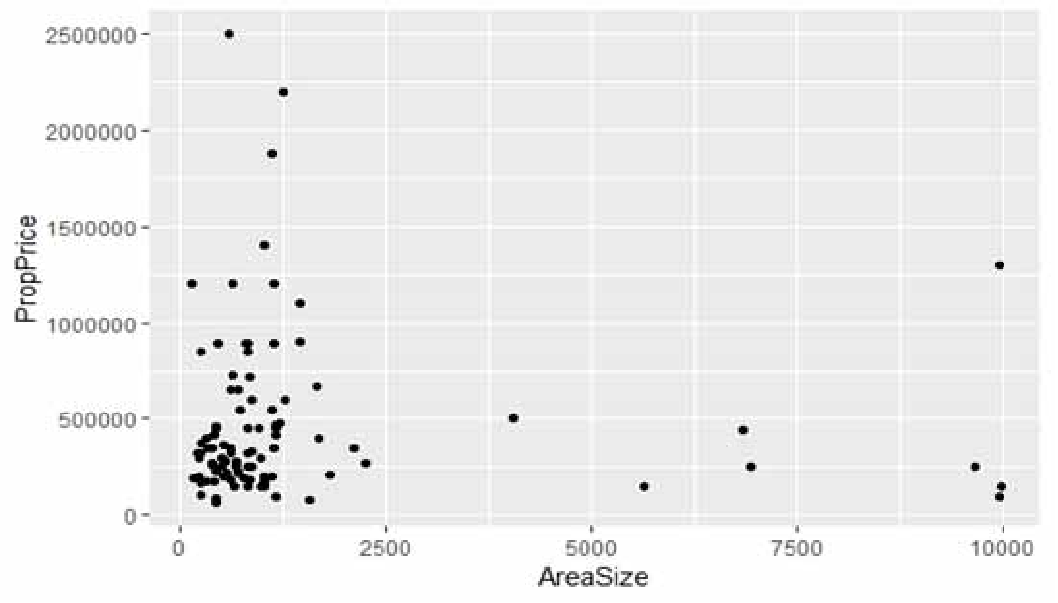
Figure 4. Variables Outline Count of Actual Attributes



Based on the variables above, for the model created for prediction, the variable which do not affect the results are eliminated to further refine and improve the accuracy. Hence, the authors create another model by not considering the variables that do not impact the price of properties (Garage, and Fully Fenced). Comparing the two models, the root mean square error is significantly lesser for the second model (24893.6) where the initial model built had a root mean square error of 29367.2. This denotes that the second linear regression model performs better, having removed the variables which do not affect the valuation prices.

This completed the first task of determining factors which impact property valuations and leads to major task of predicting prices of properties advertised on social media, and deducing any persisting relationship between the true valuations of properties to the prices that people openly sought for

Figure 5. Presence of Outliers



when selling off their properties. The authors progress by utilizing the valuation dataset with impact variables as the train data with valuation price and mined data as the test data with variables refined to contain only the impact variables, and the absence of valuation prices. Visualizing the train data (pre-processed from task one) yields the following information. As depicted in figure 4, variables outline the count of actual attributes as values present in the train dataset.

Furthermore, the dataset was inspected for outliers as depicted in Figure 5. These outliers needed to be phased to for better performance of the model in order to predict valuation by reducing the possibilities of errors.

Random Forest Model was built using the Train data set and the predicted valuation was compared with the selling price of properties advertised on the social media platform, namely, Facebook, on its user groups specializing in sale of properties focusing on mainland Fiji. Figure 6 shows the predicted valuation in comparison to the selling price posted by specific group users utilizing Facebook as a means to advertise their property. The predicted model has the root mean square error of 23046.27, which, when looking at the actual prices present in the data, is quite impressive.

If the model is to be trusted, then several noteworthy concerns are realized. Only few users are aware of the current valuation of their property, while most of the properties advertised happen to have significantly misrepresented the price of the properties been sold. Out of the many misinterpreted figured advertised, the selling price is noted to be higher than the predicted valuation. There are also a few properties which listed prices below the predicted valuation as expected. Figure 7 shows the valuation prices compared with the advertised prices by urban locations from around the mainland.

From Figure 7, it is evident that properties based in Suva are far more expensive in comparison to other towns and cities. Also, it is interesting to note that properties in Suva and Nadi happen to be selling at much higher price in comparison to their actual valuation if our model is to be trusted, whereas most of the properties in Lautoka and Nausori average on valuation prices sales. One of the factors for such observations could be the greater urban developments and migration to the major towns and cities such as Suva and Nadi.

Figure 6. Advertised Selling Prices Compared with Predicted Valuation

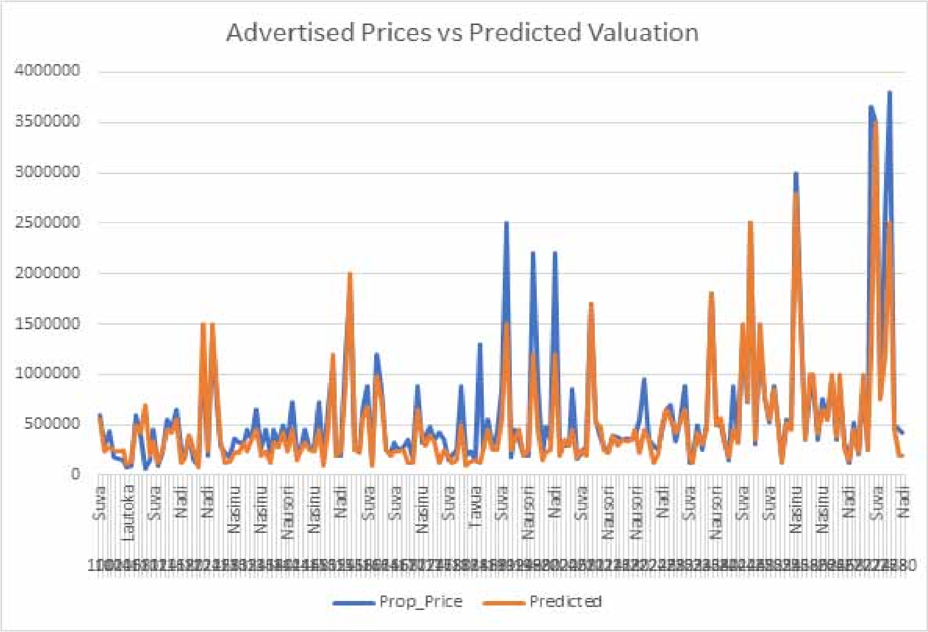
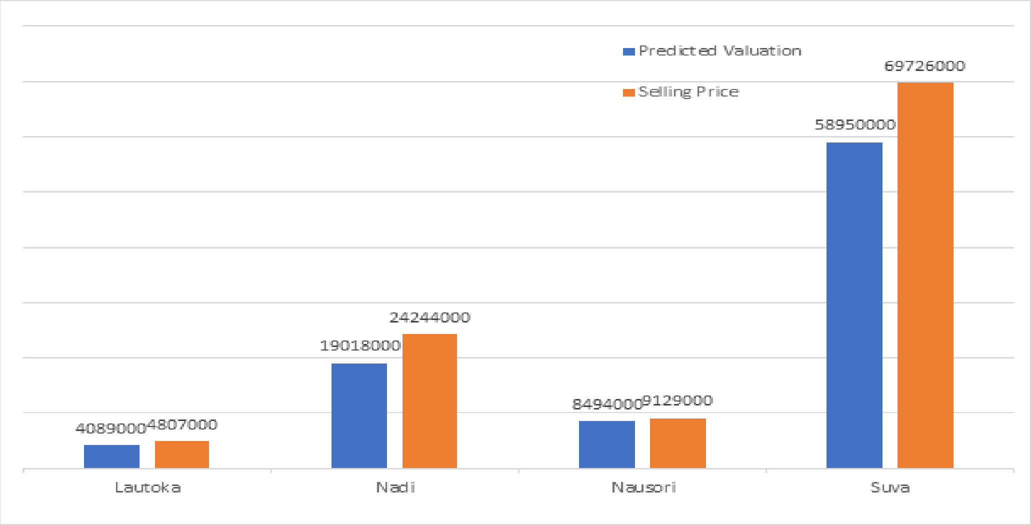


Figure 7. Predicted Valuation and Selling Price Advertised



CONCLUSION

Fiji Islands, as a republican nation has witnessed an economic boom over the recent past and is in its quest to further develop towards being the business hub of the South Pacific. As a result of such progression, a major increase in property value has expectedly been witnessed. Land tenureship, type of land, centralized location of the property and the area, in terms of the size, with other factors such type of property, number of flats and bedrooms and luxury amenities such as swimming pools are some of the factors deduced using linear regression and random forest techniques of data mining and depicted through this research, affecting the price of properties in Fiji. It is interesting to note that some of the factors, which could normally be deemed as strong variables contributing to property values, such as engineers' certificate, property title and valuation report, do not impact the determination of property price and sales much, or have very little effect. As per our research, the Residual standard error is 506600 on 231 degrees of freedom with multiple r-squared at 0.3322 when adjusted r-squared at 0.2946 displays the F-statistic as 8.838 on 13 and 231 DF, p-value at 1.289e-14.

The prices advertised on social media platforms have become a common and most searched places for property sales, eliminating and heavily impacting formal mediators such as real estate agencies, however, this paper brings to light that such means of investments and property sales are not viable as the prices being advertised are basically not in compliance with the true valuation of property prices. It has also become apparent that greater cities and urban centers around the mainland disregard the actual valuation prices and offer sale of properties on a much higher price.

Future work on this path could focus on automating the process of extracting or scraping data from web and incorporate other platforms based on the future trends. The data utilized for this experiment is that of the past six months, in order to maintain recency and validity of valuations. The predictions on valuation of prices could also be experimented using other models such as artificial neural networks utilizing time series data, which could also enable forecasting of valuation prices based on past trends and a web-based platform could be developed to enable social media users to quickly verify a good estimate of property valuation, for sellers and interested buyers alike. In most cases, these sales are often entertained due to the rate of progression and future value of these properties. Smaller towns and decentralized locations are observed to have property sale adverts below the predicted valuation and could be attributed to the slow progress due to low rate in developments in these areas. In order to predict valuation of properties advertised on social media platforms, we utilized Facebook, as platform of choice to extract data from, as it is the most common choice of social media in Fiji.

REFERENCES

Abdollahi, H., & Ebrahimi, S. (2019). Modelling and Investigating the Economy and Production Structure of Iran Public Theatre: A System Dynamics Approach. *International Journal of System Dynamics Applications, 8*(1). .10.4018/IJSDA.2019010104

Asad, A., Anwar, M. M., & Dawood, M. (2020, January 14). *The impact of neighbourhood services on land values: An estimation through the hedonic pricing model*. Retrieved March 11, 2020, from https://link.springer.com/article/10.1007/s10708-019-10127-w

Balochian, S., & Rajaee, N. (2018). Fractional-Order Optimal Control of Fractional-Order Linear Vibration Systems with Time Delay. *International Journal of System Dynamics Applications*, *7*(3), 72–93. Advance online publication. doi:10.4018/IJSDA.2018070104

Bandyopadhyay, S., & Maulik, U. (2017). Knowledge Discovery and Data Mining. *Advanced Information and Knowledge Processing Advanced Methods for Knowledge Discovery from Complex Data*, 3-42. .10.1007/1-84628-284-5_1

Banea, C., & Mihalcea, R. (2018). Possession identification in text. *Natural Language Engineering*, *24*(4), 1–22. doi:10.1017/S1351324918000062

Branko, B., Milićević, D., Marko, P., & Marošan, S. (2013, May). The use of multiple linear regression in property valuation. *Geonauka*, *1*(1), 41–45. doi:10.14438/gn.2013.06

D'Amato, M., Cvorovich, V., & Amoruso, P. (2017). Short Tab Market Comparison Approach. An Application to the Residential Real Estate Market in Bari. *Advances in Automated Valuation Modelling Studies in Systems, Decision and Control*, 401-410. 10.1007/978-3-319-49746-4_22

Doumpos, M., Papastamos, D., Andritsos, D., & Zopounidis, C. (2020). Developing Automated Valuation Models for Estimating Property Values: A Comparison of Global and Locally Weighted Approaches. *Annals of Operations Research*. https://link.springer.com/article/10.1007/s10479-020-03556-

Fayyad, U. M., Mannila, H., & Ramakrishnan, R. (2004). Editorial. *Data Mining and Knowledge Discovery*, *8*(1), 5–6. doi:10.1023/B:DAMI.0000005299.60542.d2

Galli, B. (2019). Application of System Engineering to Project Management: How They Relate and Overlap. *International Journal of System Dynamics Applications*, *8*(1), 79–93. Advance online publication. doi:10.4018/IJSDA.2019010105

Ghabi, J., Rhif, A., & Vaidyanathan, S. (2018). Discrete Time Sliding Mode Control Scheme for Nonlinear Systems with Bounded Uncertainties. *International Journal of System Dynamics Applications*, *7*(2), 15–33. Advance online publication. doi:10.4018/IJSDA.2018040102

Herrera, M., Carvajal-Prieto, L., Uriona-Maldonado, M., & Ojeda, M. (2019). Modeling the Customer Value Generation in the Industry's Supply Chain. *International Journal of System Dynamics Applications*, *8*(4), 1–13. Advance online publication. doi:10.4018/IJSDA.2019100101

Herzog, B. (2018). Valuation of Digital Platforms: Experimental Evidence for Google and Facebook. *International Journal of Financial Studies*, *6*(4), 87. doi:10.3390/ijfs6040087

Hussein, T., Ammar, M., & Hassan, M. (2017). Three Phase Induction Motor's Stator Turns Fault Analysis Based on Artificial Intelligence. *International Journal of System Dynamics Applications*, *6*(3), 1–19. doi:10.4018/IJSDA.2017070101

Irfan, R., King, C., Grages, D., Ewen, S., Khan, S., Madani, S., Kolodziej, J., Wang, L., Chen, D., Rayes, A., Tziritas, N., Xu, C.-Z., Zomaya, A. Y., Alzahrani, A. S., & Li, H. (2015). A survey on text mining in social networks. *The Knowledge Engineering Review*, *30*(02), 157–170. doi:10.1017/S0269888914000277

José-Luis, A., Emilio, C., Alfaro-Cortés, E., García, N., Gámez, M., & Larraz, B. (2020). A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems. Hindawi. doi:10.1155/2020/5287263

Krebs, F., Lubascher, B., Moers, T., Schaap, P., & Spanakis, G. (2018). Social Emotion Mining Techniques for Facebook Posts Reaction Prediction. *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*. doi:10.5220/0006656002110220

Maimon, O., & Rokach, L. (2009). Introduction to Knowledge Discovery and Data Mining. *Data Mining and Knowledge Discovery Handbook*, 1-15. 10.1007/978-0-387-09823-4_1

Milovic, B. (2012). Prediction and decision making in Health Care using Data Mining. *International Journal of Public Health Science*, *1*(2). Advance online publication. doi:10.11591/ijphs.v1i2.1380

Moysis, L., & Taher, A. (2017). New Discrete Time 2D Chaotic Maps. *International Journal of System Dynamics Applications*, *6*(1), 77–104. Advance online publication. doi:10.4018/IJSDA.2017010105

Panda, M. (2019). Software Defect Prediction Using Hybrid Distribution Base Balance Instance Selection and Radial Basis Function Classifier. *International Journal of System Dynamics Applications*, *8*(3), 53–75. Advance online publication. doi:10.4018/IJSDA.2019070103

Pilania, G., Gubernatis, J. E., Lookman, T., & Ramprasad, R. (2015). Materials Classification & Accelerated Property Predictions using. *Machine Learning*. Advance online publication. doi:10.2172/1184607

Price, C. (2017). The Statistical Basis of Valuation: The Hedonic House Price Model. *Landscape Economics*, 223-248. 10.1007/978-3-319-54873-9_12

Roshni, S., Sagayam, R., & Srinivasan, S. (2012). A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques. *International Journal of Computational Engineering Research*, *2*(5), 1443–1446.

Saravanan, D., & Chonkanathan, K. (2010). Text Data Mining: Clustering Approach. *International Journal of Power Control Signal and Computation*, *1*(4), 16.

Shukla, O., Jangid, V., Soni, G., & Kumar, R. (2019). Grey Based Decision Making for Evaluating Sustainable Performance of Indian Marble Industries. *International Journal of System Dynamics Applications*, *8*(2), 1–18. Advance online publication. doi:10.4018/IJSDA.2019040101

Stepaniuk, J. (2017). Mining Knowledge from Complex Data. *Studies in Computational Intelligence Rough – Granular Computing in Knowledge Discovery and Data Mining*, 99-110. 10.1007/978-3-540-70801-8_7

Wu, H., Liu, K., & Trappey, C. (2014). Understanding customers using Facebook Pages: Data mining user's feedback using text analysis. *Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. doi:10.1109/CSCWD.2014.6846867

Sam Goundar is an Editor-in-Chief of the International Journal of Blockchains and Cryptocurrencies (IJFC) – Inderscience Publishers, Editor-in-Chief of the International Journal of Fog Computing (IJFC) – IGI Publishers, Section Editor of the Journal of Education and Information Technologies (EAIT) – Springer and Editor-in-Chief (Emeritus) of the International Journal of Cloud Applications and Computing (IJCAC) – IGI Publishers. He is also on the Editorial Review Board of more than 20 high impact factor journals. As a researcher, apart from Blockchains, Cryptocurrencies, Fog Computing, Mobile Cloud Computing and Cloud Computing, Dr. Sam Goundar also researches in Educational Technology, MOOCs, Artificial Intelligence, ICT in Climate Change, ICT Devices in the Classroom, Using Mobile Devices in Education, e-Government, and Disaster Management. He has published on all these topics. He was a Research Fellow with the United Nations University. He is a Senior Lecturer in IS at The University of the South Pacific, Adjunct Lecturer in IS at Victoria University of Wellington and an Affiliate Professor of Information Technology at Pontificia Universidad Catolica Del Peru.

Akashdeep Bhardwaj achieved his PhD from University of Petroleum & Energy Studies (UPES), Post Graduate Diploma in Management (PGDM), Engineering graduate in Computer Science. He has worked as Head of Cyber Security Operations and currently is a Professor in a leading university in India. He has over 24 year experience working as an Enterprise Risk and Resilience and Information Security and Technology professional for various global multinationals.