2020

LONDON METROPOLITAN UNIVERSITY

CC7164 Data Mining for Business Intelligence

Prediction of Housing Market Trends in London Using Supervised Learning

Adam Adetomiwa Adeleke

***Abstract*— Analyzing market trends and cycles, this paper provides a breakdown of potential investment opportunities as a result of trend speculation based on factors which price of a residential property are responsive to and supplemented by past data. A thorough overview of impactful factors to the risk involved is explored with the acknowledgment of scenarios yielding a highly probable-profitable return.**

Keywords: Price, Rent, Tax, Earnings, Return-on-Investment, Postcode, Ward, Borough, Decision Tree, C4.5, Logistic Regression, Neural Network, GAMLSS, Multilayer Perception, ROC, Supervised Learning, Data Mining, CRISP-DM, Centile Fan, Probability Distribution

I. BUSINESS AND DATA UNDERSTANDING

The housing market can be characterized by regular market conditions, meaning the influence of supply and demand are effective on price. Driving forces behind market price include fundamental effects of supply such as cost of construction and developable land as well as effects of demand which includes (not limited to) the availability of credit [1].

This market's attribute is unique due to the nature of its commodity. The demand for the good is as a result of the good itself rather than for its consumption or ability to access other types of goods/services. The effects of the intrinsic value of its good are the inability of fundamentals to solely determine its price. Therefore, further limitations to direct effects of demand and supply include:

- Speculative use by investors with supply dampening once units are left vacant.
- Scarcity of developable land.
- Time taken for construction of new properties coupled with further issues involving planning restrictions.
- The increased cost of doing business due to wage increase caused by factoring housing cost for employees as well as transport costs.

In terms of the general market trend in the last decade, there has been a successive year-on-year increase in property value. This makes the market very lucrative for investors. However, the subjection of the commodity to factors besides fundamentals accommodates volatility especially in scenarios of price correction/equilibrium, hence increasing risk.

A. Overview

The approach is taken towards the analysis and then the prediction is done from the perspective of an investor. Speculation of market trends using factors (rather the change of factors in certain instances) that determine the valuation. The prediction of trends will be done on an individual basis according to property types which include:

- Flats
- Semi-Detached
- Detached
- Terraced

Specification of the property type reduces the haphazard nature of market classification by narrowing down clusters for unique characteristics for factors relating to each property type.

B. Background

The variables used to reflect the yearly characteristic of each datapoint with those datapoints being according to the postcode sector in London. Some variables are not specific characteristics of the postcode sector but do relate to the datapoint if that postcode sector is within a given ward or borough. The following are the variables used for trend prediction:

- Price Paid: This variable represents the average amount paid in a transaction for specific property types within a postcode sector. The length of time used for calculating the average price is one year and the ***response variable*** is extracted by calculating the year-on-year price change to prior price ratio (percentage change). Furthermore, the ratio of the postcode sector price to the borough price is set as an ***explanatory variable***. The price ratio is not an extract of the response variable as all explanatory variables are factors relating to the present time (*Time(x)*) and the former characterizes the future (*Time(x+1)*). The data is sourced from HM Land Registry [2].
- Sales: This represents the number of transactions in a calendar year. The data available include postcode sector sales and borough sales. Sales represent demand with the net change in year-on-year transactions as a ratio to prior sales being the variable(s) of focus. The data is sourced from HM Land Registry [2].
- Mortgage Interest Rate – The only generic variable in the model renders data for an assessment of credit. The ratio of the change in percentage point of mortgage interest rates to prior rates essentially highlights the year-on-year attractiveness of credit in London. The source is the Bank of England with 2011 to 2015 data being the standard variable mortgage for households [3] and 2016 to 2020 data being the standard variable mortgage for individual trusts [4].
- Population Density: This highlights congestion within a ward. Already serving as a ratio to indicate the number of individuals within a hectare, the year-on-year percentage change of this factor will represent an explanatory variable. The data is sourced from the Greater London Authority which includes GLA estimates, 2011 census, and ONS mid-year estimates for the population [5].
- Mortgage Lending: This represents the value of outstanding mortgages in a specified postcode sector. The general idea behind this variable is to render an assessment of credit in a specified postcode sector. Year-on-year changes as a ratio to a prior outstanding mortgage will serve as an explanatory variable for the

model. The data is sourced from UK Finance mortgage lending from 2013 to 2019 [6].

- Transport Assessment: The Public Transport Accessibility (PTAL) data highlights the quality of transport in various wards. The higher the PTAL value of a ward, the better the transport quality. The data is sourced from Transport for London [7].
- Housing Stock: This represents the current supply in the market. The data source is from the Greater London Authority [8]
- Additional Dwellings: This represents the increase in housing supply. The explanatory variable will be a ratio of additional dwellings to the housing stock for a given year. The data will represent borough characteristics with the intent of including an aspect of supply to the model. Data is sourced from the Greater London Authority [8]
- Earnings by workplace: A factor influences the housing cost through attractiveness for residence by employees. According to the limitations on demand and supply mentioned, a significant increase in the change in earnings ought to be accompanied by an increase in demand. The effects of workplace earnings is expected to be further supplemented by transport quality in the model. The variable to be used is the ratio of the change in average earnings to the prior earnings by workplace which is according to each borough in London. Data is sourced from the Office of National Statistics [9].
- Earnings by Residence: In terms of the model, this variable partly delivers a rationale for the standard of living within each specified borough hence translating to varying housing cost. The explanatory variable will be the ratio of change in earnings to prior earnings by residence. Data is sourced from the Office of National Statistics [10]
- Employment Rate: Rate of employed individuals residing in a borough within the ages of 16 – 64. The year on year change in percentage points will account for a variable hence the rate of employment change. Data is sourced from the Office of National Statistics [11]
- Council Tax: The median council tax band of a borough (from band A – H) will be used to estimate the yearly percent change which is the variable to be considered. Data is sourced from the Ministry of Housing, Communities, and Local Governments [12].
- Rent – Here, rent will represent the estimated cost of housing as a percentage of the house price to result in the rental yield. Data is sourced from the Valuation Office Agency [13]

C. Business Objective

The *response variable* titled **Forecast** is then transformed from an interval type variable to a Binary variable.

The Binary variable is then labeled as **ROI** (Return on Investment) which is split at a mid-point of the Forecast variable with a value zero (0.00) where greater than the mid-point is labeled as **PROFIT** and less than the mid-point is labeled as **LOSS**.

Variable selection varies based on multiple factors which include:

- Type of Model: Criteria for variable selection becomes much stricter incases of the decision tree as opposed to neural networks. This reason is due to the type of explanatory variables used which are all intervals and best handled with neural networks.
- Trend Characteristics: The ability of a variable to possess a definite distinction between the Binary datapoints of the response variable will adversely affect its selection. Furthermore, trends observed by the explanatory variables against Forecast will be a significant criterion for variable selection.

The selection of viable postcode sectors for the result is done using classification probabilities equal to/greater than 90%. The interpretation of the ROI to Forecast is done using probability distributions from test samples for 10 to 5 blocks (depending on the sample size) of the 90%-100% classification probability range for PROFIT.

II. DATA PREPARATION AND EXPLORATION

A. Data Selection and pre-processing

The inclusion or exclusion of datasets proposed is done based on time available. Excluded dataset include Age, Crime Rate, and Job Density. The majority of datasets range between 2011 to 2020 with Mortgage Lending being the only exception. Datasets are segmented into portions that are according to scale referring to the area covered in London. These segments are as follow:

- Postcode Sector: Price Paid, Sales, and Mortgage Lending.
- Ward: Population Density and Transport Assessment.

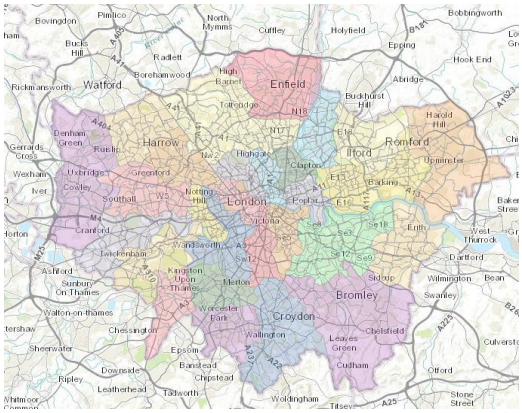


Figure 1. Map of London Boroughs and Wards [14]

- Borough: Housing Stock, Additional Dwellings, Dwellings per Hectare, Earnings (Residence and Workplace), Employment Rate, Council Tax, Rent, and Price Paid.

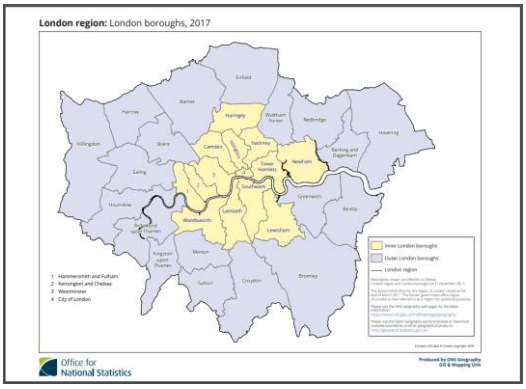


Figure 2. Map of London Boroughs [15]

- London: Mortgage Interest Rate

The range of historical data used is from 2014 to 2019. The processing of multiple datasets into a single dataset is done using a postcode directory. The postcode directory contains data used for referencing specific datapoints to a generic ID before further processing. Essential IDs extracted from multiple data sets include; Borough Code, Ward Code, and Postcode Sector. The Postcode District is extracted to ease referencing to specific wards.

TABLE I

POSTCODE DIRECTORY

Pc District	Pc sector	Borough Code	Borough	Ward Code	Ward
BR1	BR1 1	E09000006	Bromley	E05000109	Bromley Town
CR2	CR2 0	E09000008	Croydon	E05011478	Sanderstead
EC3V	EC3V 1	E09000001	City of London	E05009307	Lime Street
HA2	HA2 0	E09000015	Harrow	E05000289	Harrow on the Hill

The ID finally used for data points consisted of a combination of a datapoint's Postcode Sector, Ward Code, Year, and Type. Property Types are as follow:

1. Semi-Detached
2. Detached
3. Terraced
4. Flats

Below is an example of a data point with an ID for a Detached property within BR1 1 in Bromley Town with data for 2011.

Postcode Sector	Ward Code	Year	Type		
↓	↓	↓	↓	→	BR1 1E0500010920112
BR1 1	E05000109	2011	2		

Figure 3. Illustration of the final dataset ID

TABLE 2

HOUSING PREDICTION DATASET

ID	Year	Var 1		Var 14	Price Change	Forecast	ROI
BR1 3E0500012520142	2014	1.1618		0	0.05392714	0.426598	PROFIT
BR1 3E0500012520152	2015	1.5197	...	0.00281	0.42659782	-0.07847	LOSS
BR1 3E0500012520162	2016	1.3455		-0.06443	-0.0784669	-0.1684	LOSS
BR1 3E0500012520172	2017	1.0659		0.07186	-0.1684042	0.135177	PROFIT

The above table partially illustrates the final dataset but mainly highlights how the response variable (ROI) is attained. The forecast variable is the price change for a time Year(x+1) afterward the ROI is then classified.

The following is the summary of variables included in the model:

- ROI (Response Variable) – A binary variable categorizing the Forecast variable into PROFIT or LOSS.
- COSTR – Cost ratio highlights the ratio of the price paid in a postcode sector to the overall borough.
- PVOR – Postcode volume ratio illustrates the year-on-year net change in sales within a postcode sector.
- BVOR – Borough volume ratio illustrates the year-on-year net change in sales within a Borough.
- MCAP – A variable for the net change in year-on-year outstanding mortgage value for a postcode sector.
- RYDS – This is the ratio of the average rent to the price of a property.
- PDES – The net change of the populating density (population per hectare).
- TASS – Transport assessment (PTAL) per ward.
- SADD – A ratio of the additional dwellings to the housing stock.
- DWELLD – The rate of change in the number of dwellings per hectare.
- REARN – The rate of change in median salary earning by residence within a borough.
- WEARN - The rate of change in median salary earning by employees within a borough.
- EMPR- The net change in the employment rate.
- MTAX – The rate of change of median council tax.
- MRGI – The rate of change of mortgage interest.

B. Data measurement setting and transforming

The majority of the variables used in this model are ratios. However, variable classification is according to the following:

- Net Change Ratio: These variables can be easily understood as the percentage change of a parameter. Rather than use a multiple of 100, the range of these ratios is from zero to one (0,1). An interpretation of the model can be described to be the prediction of returns based on the changes in these factors. These variables include PVOR, BVOR, MCAP, DWELLD, REARN, WEARN, MTAX, and MRGI. The distribution type for variables mentioned is mixed with all variable types being intervals.

- Generic Ratios: This includes COSTR and SADD. The distribution type is continuous with limits from zero to infinity (0, ∞).
- Assessment: TASS is the only variable that remains unchanged over the time range of 2014 to 2019.

The dataset is divided into four separate types according to property types to build different variables independent of the type of property. Although parameters for modeling may vary, variable exploration may not (slight variations may apply). Furthermore, the analysis of variables will be for Flat data.

Skewness in data tends to reduce the quality of the model result hence the transformation of variables to reduce that distributions parameter is essential to produce reliable results.

The first variable transformed is COSTR with the initial range of [0.26, 6.47] and a median value of 0.94. The logarithmic transform of the variable produces a range of [-1.35,1.86] and a median of -0.08 which drastically reduces skewness.

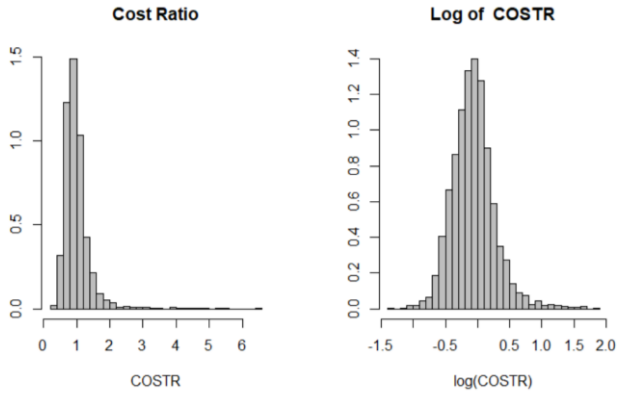


Figure 4. Distribution of the Cost Ratio variable and its Logarithmic Transform

The distribution of the rental yield has no issues with skewness. However, the range of the variable poses an issue with models that require weightings. The identity for Rental Yields has a range of [0.00026, 0.00918] and the transformed exponential renders a range of [1.00026, 1.009266]

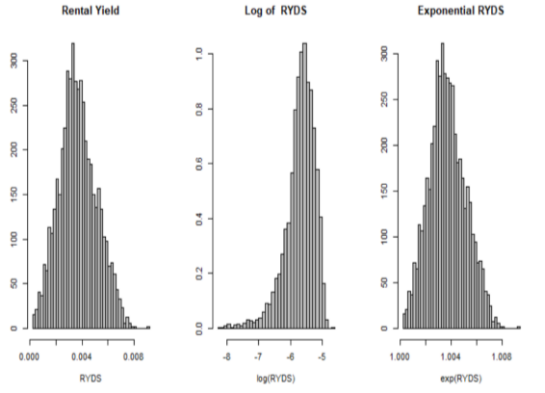


Figure 5. Distribution of the Rental Yield Variable illustrating the identity, logarithmic and exponential functions.

Transport Assessment has an initial range of [0.37, 83.45] and a median value of 11.06. The logarithmic transform of the variable produces a range of [-1.01, 4.42] and a median of 2.40 which highlights reduced skewness.

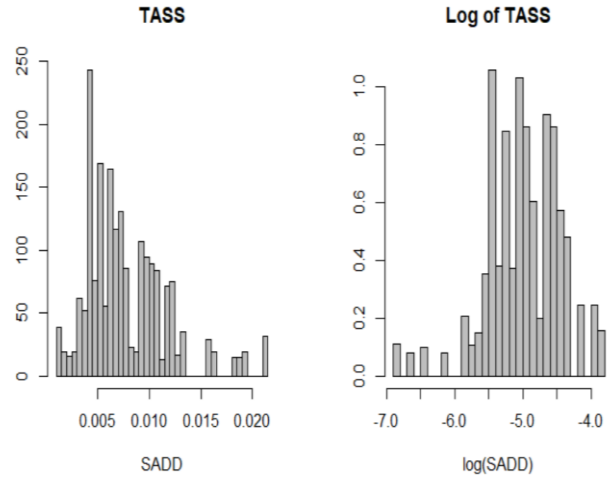


Figure 6. Distribution of the Transport Assessment Variable and its Logarithmic transform.

Similar logarithmic transforms are done for the DWELLD, SADD, and PDES. The dwelling density variable has an identity range of [0.0099, 0.8366] with a median of 0.1936. The log transform of DWELLD yields a range of [-4.61, -0.17] with a median of -1.64. The SADD variable initially has a range of [0.001, 0.021] with a median 0.006 hence the use of a logarithmic transform reduces skewness as seen with a range of [-6.82, -3.84] to -4.98 as the median.

C. Initial Data Exploring

Exploration of the dataset begins with narrowing the variables which significantly affect the response variable. Exploring the *sequential* R square values to drop variables is a reliable means for selection especially for classification models. However, due to the complexity of the variable types, selections will not be limited to that aspect. As seen in figure 7, only 5 of the 13 variables are selected which can limit adversely limit predictions therefore further analysis of variable selection is done to present enough data for the neural network.

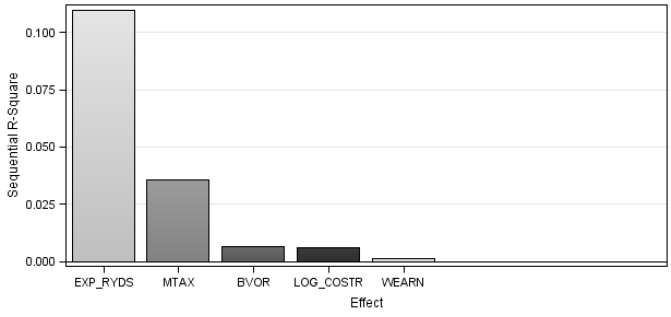


Figure 7. Sequential R-Square values of explanatory variables

An overview of R-Square values shows the significance of certain variables which were dropped while using the sequential. The logarithmic transform of DWELLD and TASS are then considered for inclusion.

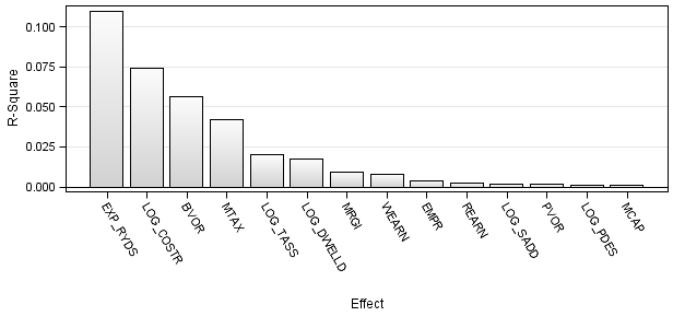


Figure 8. R-Square values of explanatory variables

The mean deviation observes how the explanatory variables show a clear opposing trend in the response variable. Using this rationale MRGI is selected for the neural network model which increases the variables used to 8.

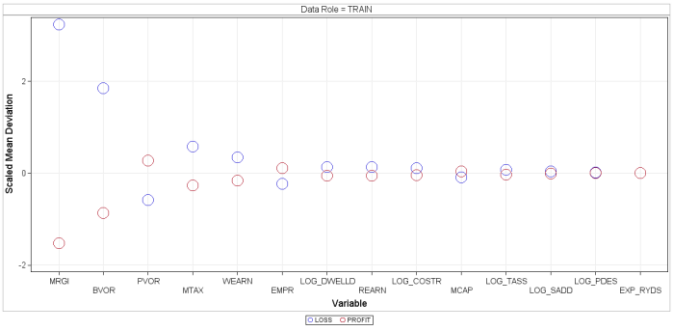


Figure 9. Scaled Mean Deviation of explanatory variables

Further analysis done will highlight the trends of the explanatory variables to examine proportionality. Inverse or direct proportionality as well as variance would be analyzed using centile estimates and visualized using centile fans.

The cost ratio as seen in fig 11. shows a dense range around [-0.5, 0.5] for FORC within the majority of the cost ration distribution between [-1, 1].

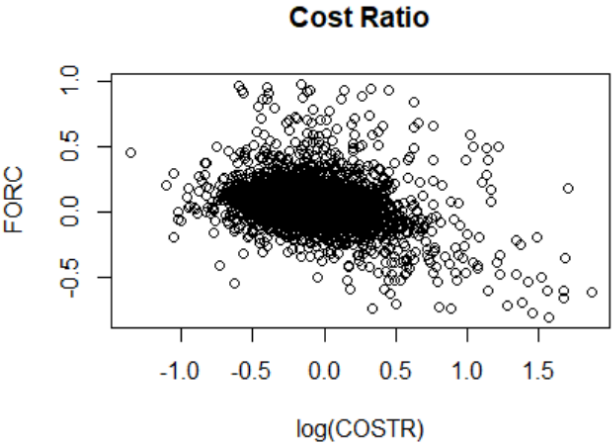


Figure 11. Forecast against Cost Ratio

The Box plot for the binary response variable shows a higher median for LOSS when compared to PROFIT. The width of the interquartile range appears to be much wider for LOSS as seen in figure 12.

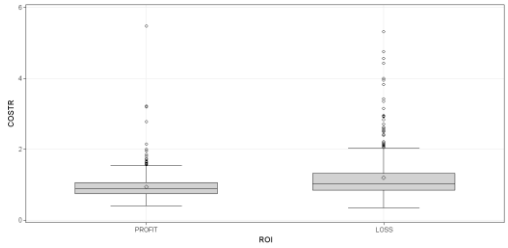


figure 12. Box Plot of the response variable ROI against the explanatory variable COSTR

Modeling the FORC against COSTR using scale and location parameters using the skew t type 1 distribution shows two factors. An inversely proportional relationship as observed in the decline median value of FORC shows a higher likelihood for profit with a smaller cost ratio. However, the variance increase with the increase in COSTR meaning the model has to manage higher volatility with a higher cost ratio. In conclusion, the lower the property price relative to the borough price, the higher the probability of ROI.

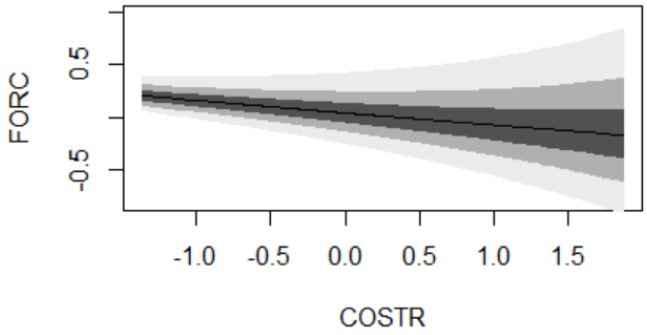
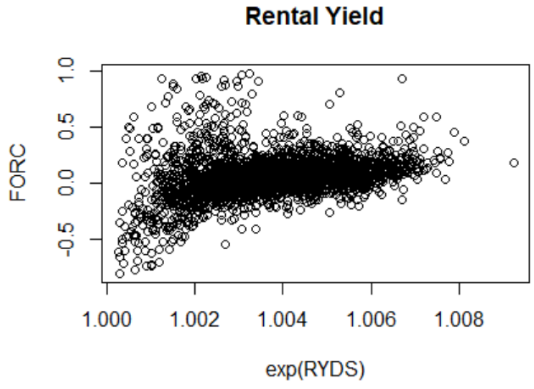


Figure 13. Centile Fan illustrating Forecast against Cost Ratio

Rental yield shows a relatively distinct relationship with the explanatory variable. This realization supplements the R2 and sequential R2 analysis of the variable.



.Figure 14. Forecast against Rental Yields

Further analysis of the RYDS variable is done by fitting a model to FORC. The skew t type 5 distribution is used while modeling according to the location and scale. An increase in the

fitted median value is observed with an increase in FORC. A corresponding decrease in variance is seen with higher RYDS.

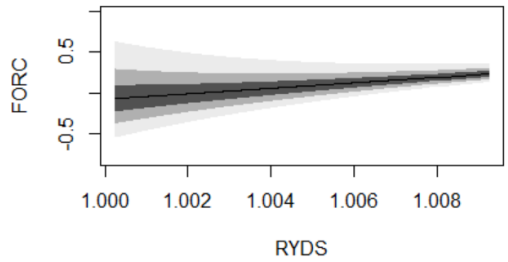


Figure 15. Centile Fan illustrating Forecast Against Rental Yields

A higher RYDS value is observed for classified PROFIT data points as opposed to LOSS. However, the interquartile ranges show no difference which furthers the narrative of the scaled mean deviation when the RYDS variable shows the least deviation amongst all explanatory variables.

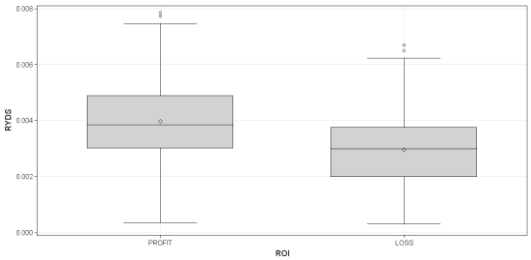


Figure 16. Box Plot of the response variable ROI against the explanatory variable RYDS

Exploring the MTAX variable shows a similarity to the remaining explanatory variables. No certain trend is observed which implies the need for more data to consolidate a relationship with the response variables. Therefore, these variables are reliable, if they are not solely used to create a rationale for the target variable.

The median council tax is a Borough type variable which means the data is not expected to be as voluminous as COSTR or RYDS hence the observation of vertical clusters in fig 17.

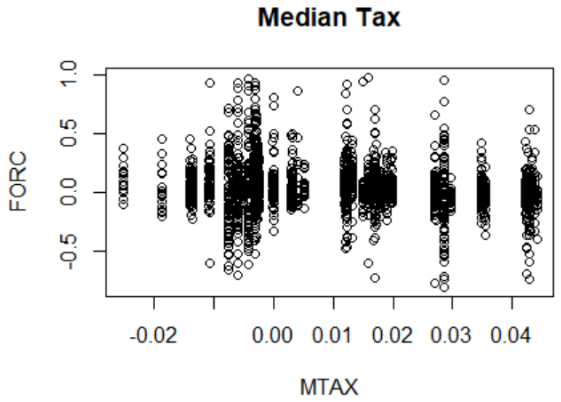


Figure 17. Forecast against Median Council Tax

Only a slight trend is observed using the fitted model. Based on a Johnson's SU repar. distribution, the median seems to decline with an increase in MTAX followed within a corresponding decline in variance.

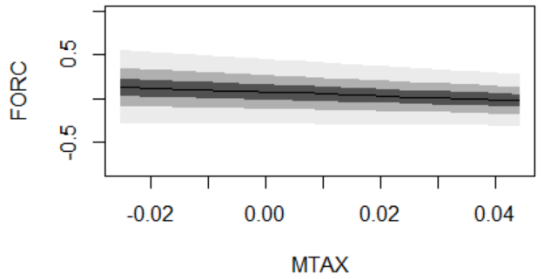


figure 18. Centile Fan illustrating Forecast Against the net change in Median Council Tax

Relative to the classification of forecast, no apparent distinction of values is seen for both response datapoints as observed in figure 19.

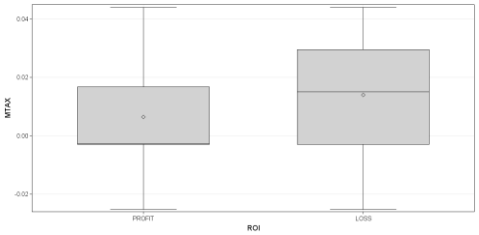


Figure19. Box Plot of the response variable ROI against the explanatory variable MTAX

All initial data analysis is done for the flat dataset which means subtle variations may apply but similar trends are observed nonetheless. Hence, a 2:1 ratio for PROFIT to LOSS is expected for all datasets just as seen with the flats in figure 20.

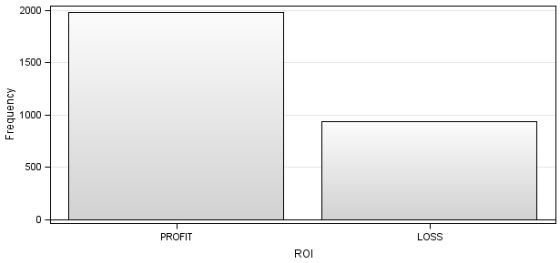


Figure 20. Frequency of ROI

III. MODELLING

A. Selecting Modelling Techniques

The first step taken towards modeling is the partitioning of the dataset. 50% of the dataset is allocated towards training the model with another 25% used to validate the model created and a further 25% used to test the model. 2014 to 2017 data is used for training, validation, and test while 2018 and 2019 are used for scoring. 2018 out of sample score dataset will provide a distribution of return as results are currently available for those predictions.

Variable transformation follows partitioning with feedback fed back to the model once suitable transformations are attained. Variable selection filters out the significant variables for both decision tree and logistic regression models while additional variables are included for neural networks.

1) Decision Tree

The ***Decision Tree*** model used is an adaptation of the C4.5 algorithm [16]. As a result, the impurity selection criteria is done using the entropy impurity index. The purpose of splitting is to ensure as reduction of impurity and calculations are done using the information theory:

$$E = -\sum_{i=1}^{2} p_i \log_2(p_i) = -\frac{\sum_{i=1}^{2} p_i \log_e(p_i)}{\log_e(2)}$$

Equation 1. Entropy Index Calculation where p_i represents probabilities in the parent node

$$E_0 = -p_1 \log_2 p_1 - p_2 \log_2 p_2$$

Equation 2. Binary Split Calculation for Entropy Impurity Index with p_1 and p_2

$$\Delta E = I_0 - \left[\frac{n_1 E_1}{n_0} + \frac{n_2 E_2}{n_0}\right]$$

Equation 3. Calculation of the change in Entropy Impurity ΔE.

Consequently, the use of this method is to ensure the same amount of impurity in the parent node I_0 (with the number of cases n_0) exists in the child nodes (E_1 and E_2 with given n cases). The maximum number of branches is set to 2 (for binary splits) with a maximum depth is set to 5 which limits the number of generations to that value. The Minimum amount of observations allowed for further splitting (Split Size) is set at 20 cases which is the same as the maximum number of cases a node can possess while the number of rules for Entropy reduction is set at 2. The subtree construction method is based on Assessment which is measured based on Decisions. The decision tree assessment is based on the tree that has the largest average PROFIT and smallest average LOSS.

2) Logistic Regression

The ***Logistic Regression*** model operates based on a stepwise selection model. This implies significant effects (variables) that impact the model are chosen then reviewed to observe those relatively insignificant which are then dropped. The process of stepwise selection is done until an Entry and Stay significance level are met. In the case of this model, those values are 0.01 each. GLM (Generalized Linear Model) encoding is used to estimate the difference between each parameter level and then reference them. Furthermore, the intercept is suppressed which means only effects are strictly variables selected.

3) Neural Networks

The ***Neural Networks*** model makes use of a multilayer perception architecture which means it was composed of an input layer, hidden layer, and output layer with a direct connection between the input and output. No specific activation link is used meaning a default setting is placed which means the activation function is dependent on the target variable measurement. Options range from Sigmoid, Tanh, and SoftMax transforms but is ultimately down to the distribution of the target.

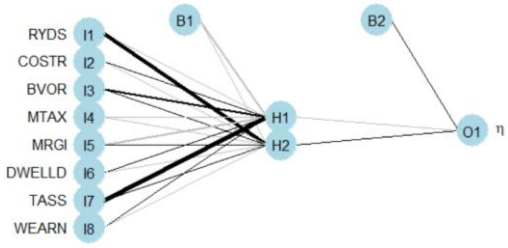


Figure 20. Neural Network Illustration

In this case, a bias the hidden layer is included while a target bias is excluded from the neural network while the number of hidden layers used is 2. The model classification criterion used is Misclassification meaning the node with the smallest classification rate is selected and the model is expected to complete training within 50 iterations.

B. Design Model Flow Diagram

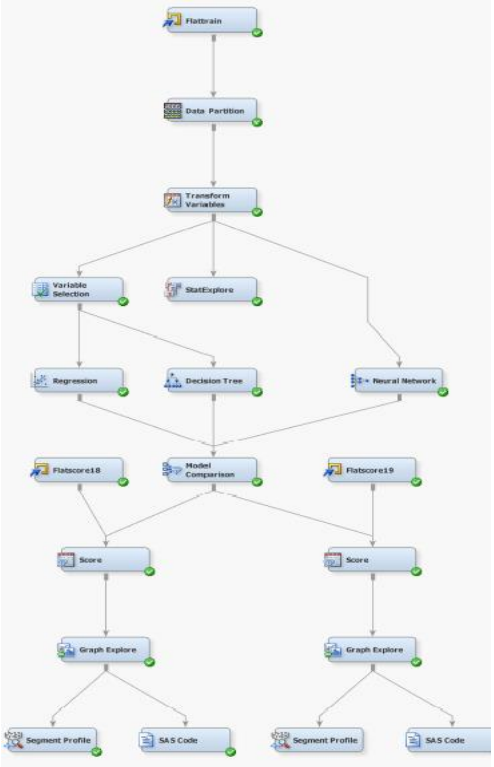


Figure 21. Model Flow Diagram

C. Models Execution and outcomes

1) Decision Tree

The decision treemap as illustrated below shows the splits of successive child nodes from the parent node. A wider width of rectangles corresponds with a larger number of cases within a node while a darker shade of grey highlights a higher probability for PROFIT.

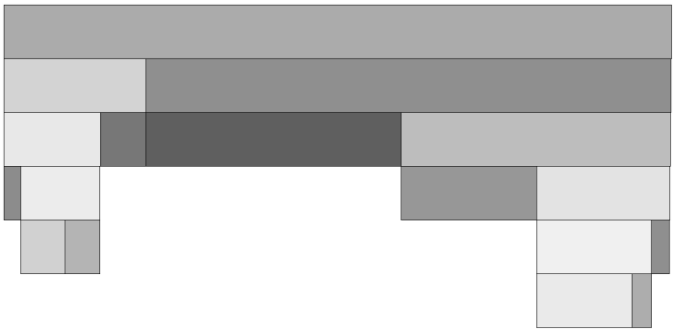


Figure 22. Decision Tree Map

The leaf statistics provides an outlook of the order of success in determining the level of PROFIT within the model. As shown in figure 23, the nodes are represented in a bar chart and arranged in order of train probability along with validation probability. The best-case node yields a probability of 87% for PROFIT.

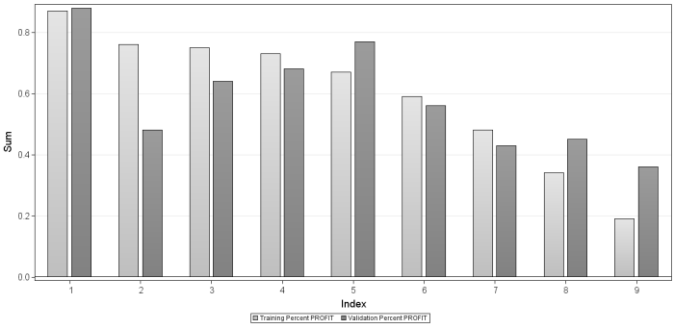


Figure 23. Decision Tree Leaf Statistic

The train, validation, and test misclassification rates are 25.39%, 28.06%, and 28.6%. The accuracy of the model ranges between 74.6% to 71.35% with PROFIT classification as high as 85.42% which dwarfs the LOSS figure of 51.60%.

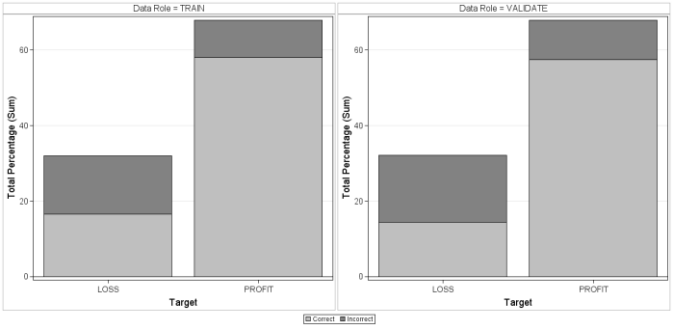


Figure 24. Decision Tree Classification Chart

2) Logistic Regression

The outcome of the stepwise selection is evaluated through the effects plot. As seen in figure 25, the magnitude and relationship between the variables as outcomes of logistic regression. The Median Council Tax (1) variable has the largest effect on the output probability for profit, at the same having an inverse relationship with the output. The magnitude of variables BVOR (4), COSTR (2), and RYDS (3) are equal but only COSTR has the same relationship as MTAX.

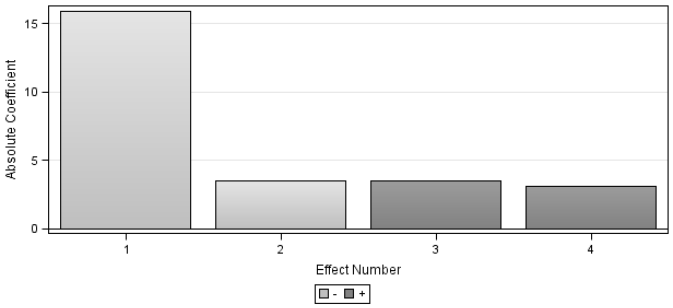


Figure 25. Logistic Regression Effects Plot

Below is a plot illustrating the absolute coefficient values of the regression model variables during the selection process. This demonstrates the procedure which occurred during the stepwise selection where RYDS was the first and only significant variable selected but the fourth (and final) step shows a similar graphic to the effects plot (relative to weightings) in figure 25.

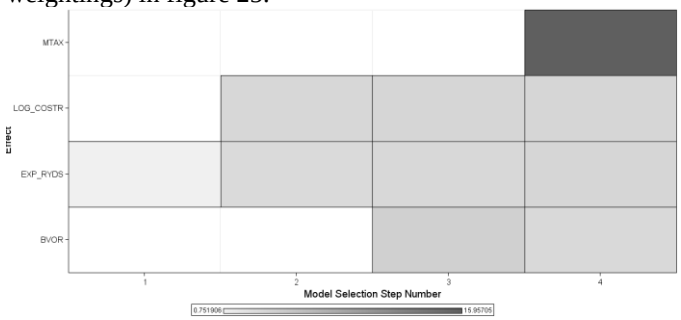


Figure 26. Logistic Regression Estimate Selection Plot

The train, validation, and test misclassification rates are 28.35%, 29.02%, and 29.2%. The accuracy of the model ranges between 71.65% to 70.8% with PROFIT classification as high as 91.80% which dwarfs the LOSS figure of 28.8%.

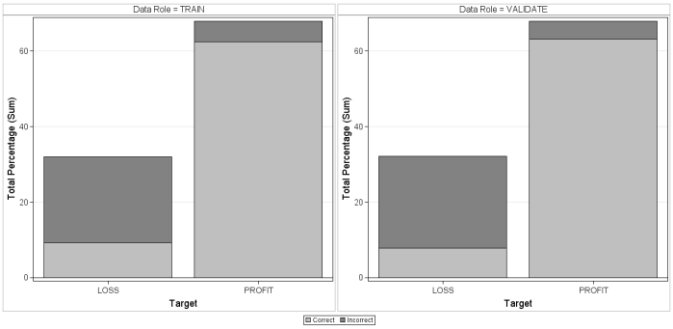


Figure 27. Logistic Regression Classification Chart

3) Neural Networks

The weights of the neural networks highlight the effects of the hidden and input layers on the output. The hidden layer H2 has a significantly more impact on the output in the order of magnitude relative to H1. Both H1 and H2 have negative weights on the output with MRGI and RYDS possessing the most significant weightings for both hidden layers.

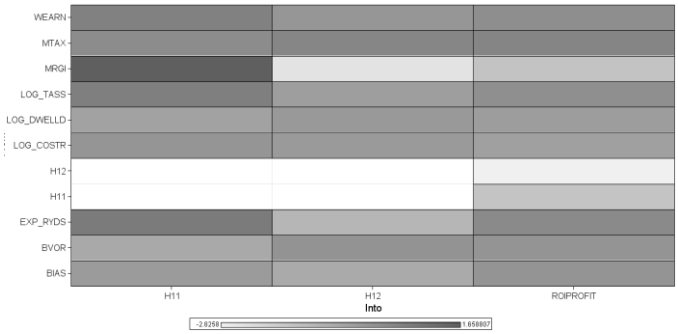


Figure 28. Neural Network Variable Weights (Final)

The number of iterations set for model training is 50 but completed within 23 which is satisfactory considering the model did not exhaust all 50 iterations before completion. The AIC is a statistic used to measure model performance; best compared with other models, the purpose here is to illustrate the reduction in AIC value from 1503 at the beginning of training to 1494 at 23 iterations.

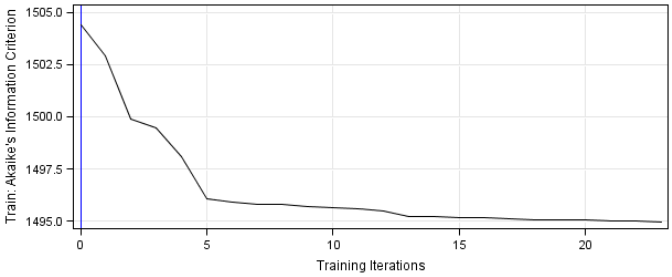


Figure 29. The Neural Network Akaike's Information Criterion (AIC) against Iterations

The train, validation, and test misclassification rates are 26.22%, 26.68%, and 29.33%. The accuracy of the model ranges between 73.78% to 70.67% with PROFIT classification as high as 86.5% which dwarfs the LOSS figure of 46.66%.

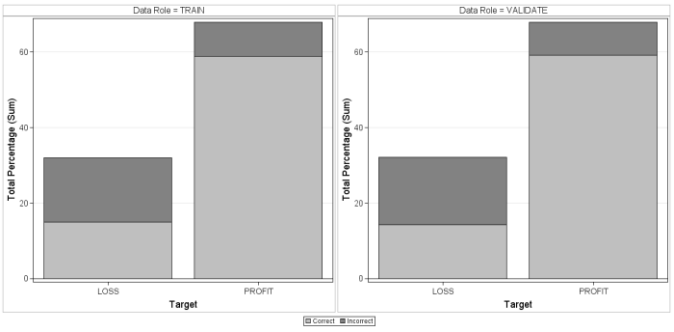


Figure 30 Neural Network Classification Chart

IV. EVALUATION AND RESULTS

A. *Models Comparison*

Selecting the optimal model is a crucial part of the data mining process. The accuracy of a model should be best perceived over segments of result rather than overall misclassification rates. A model's performance can easily be misinterpreted if various sections of its output are not properly scrutinized.

The model selection criterion is based on the ***Lift.*** This is a ratio of the response rate within a segment of a data set to the overall response rate. These segments are referred to as deciles which represent 10% divisions of the posterior probability. The response rate refers to the ratio of cases captured to the number of cases within a decile.

As seen below, there is no clear-cut model superiority over all 10 Decile blocks possess varying performance over different depths.

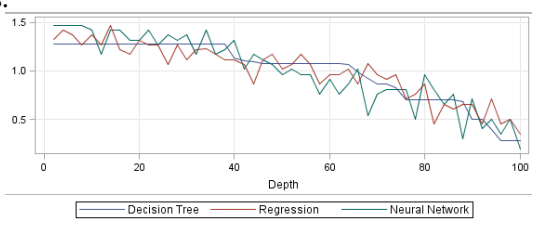


Figure 31. Lift Value against Decile Blocks for Decision Tree, Neural Networks and Logistic Regression

The cumulative Lift easily filters the noise from the Lift and better visualizes model performance. The most adequate model is one that attains the highest cumulative lift and, in this case, the Neural Network Model suffices.

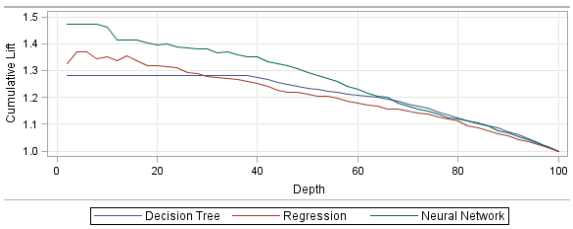


Figure 32. Cumulative Lift for Decision Tree, Neural Networks and Logistic Regression

The Receiver Operating characteristic (ROC) is a classification which highlights the gradient of Sensitivity against Specificity. Sensitivity illustrates the probability of correct classification of PROFIT relative to all classifications while Sensitivity highlights LOSS. An ideal model possesses higher sensitivity and specificity hence Neural Networks.

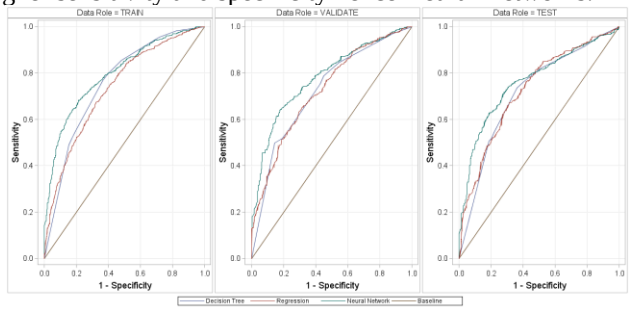


Figure 33. ROC Chart

B. *Results Explanation and Evaluation*

The probability distribution below shows the likelihood of classification from the Neural Network Model. Recalling the 2:1 ratio of PROFIT to LOSS, it becomes fitting for the majority of classification to have higher probabilities for profit between 85% to 100%.

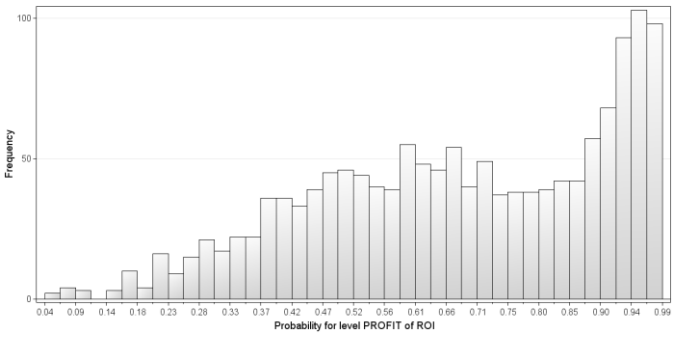


Figure 34. Probability Level PROFIT of ROI

Reviewing the levels of probability against the predicted Forecast shows a significant decline in levels of LOSS with a corresponding increase in probability. At a range of over 95%, model accuracy proves to be exceptionally accurate at classifying ROI with a much denser level of PROFIT observed. A facto worth noting is the relationship between the Forecast variable value and probability which hardly shows any correlation between those two variables.

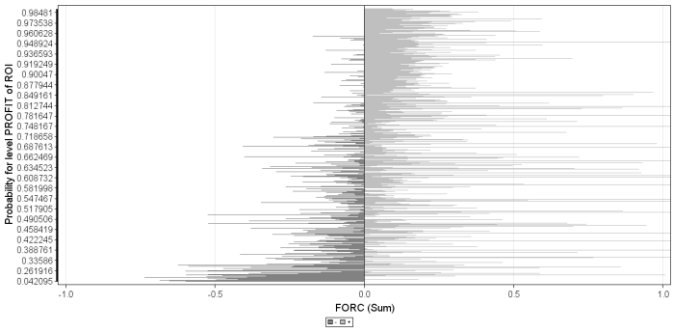


Figure 35. Probability Level PROFIT of ROI against Forecast

Further analysis dissects the Forecast variable which represents the interval version of the Binary response variable. Data extracted from the Probability level for PROFIT over 85% represent the suitable range for satisfactory Postcode sectors worth investing. The model's performance is assessed using test data partitioned from the main data set and out of sample data (Score Dataset) for 2018 to make 2019 predictions.

An overlook of the return distribution from in sample and out of sample data highlights model limitations as well as areas of advantage.

1) Detached

Test data shows a return distribution with an attractive right tail which illustrates an increase in probability for higher returns. The accuracy of the model based on binary classification is 76.47% with a median return value of 14.26%. Based on the fitted distribution, a return of over 7.17% is attained in over 65% of cases recorded. An overview of risk shows a probability for LOSS between 23.52% and 20.4% according to model accuracy and the fitted distribution respectively. The distribution then shows a probability of LOSS exceeding -5% at 12.71%. However, as stated earlier concerning the right tail, PROFIT describes a bulk of the distribution with the chance of returns over 5% at 67.78% and the probability of returns exceeding 10% at 58.46%.

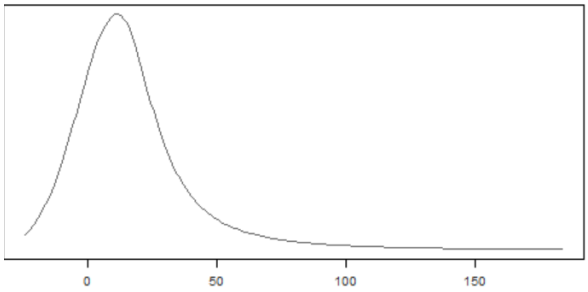


Figure 36. ROI Probability Distribution for Detached Test

The score provides less optimistic data with a median return of 4.59%. Accuracy drops significantly to 63.64% with the risk of a loss over 5% and as high as 29.11%. However, despite a reduction in quality, right tail characteristics remain with a probability of profit over 5% at 51.86%. Although less impressive, returns over -1.91% is expected in over 65% of cases with chances of greater than 10% return at 43.08%.

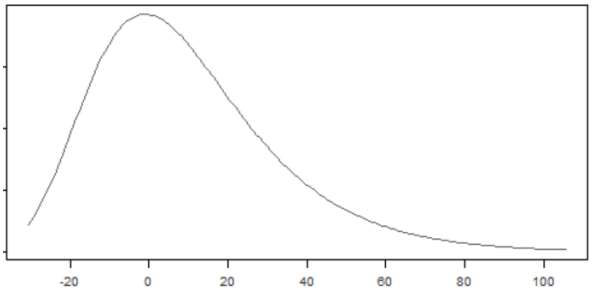


Figure 37. ROI Probability Distribution for Detached Score

2) Semi-Detached

Based on test data, the posterior probability for PROFIT ranges between 92.17% and 91.93%. A Median return of 12.16% and a return greater than 9.06% places the success rate of the model into a proper perspective. Risk shows a probability of returns over -5% at 3.08% which is minuscule compared to 80.55% and 6.75% for returns exceeding 5% and 10% respectively.

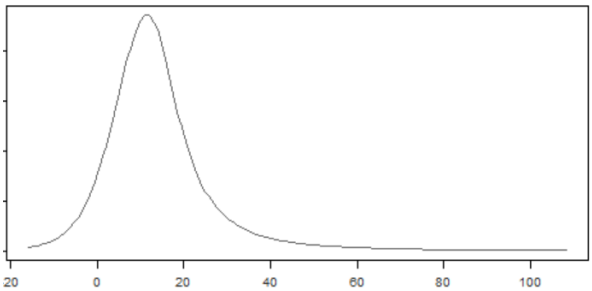


Figure 38. ROI Probability Distribution for Semi-Detached Test

Another score performance which shows discrepancies from the test but can be primarily attributed to an inadequate number of cases. The median return value is 7.24% with an accuracy of 69.23% and a probable loss exceeding -5% at 32.31%. The probability of exceeding 10% in return equals than -5% for loss and the poor performance is further complemented by only half of cases exceeding 5% in return.

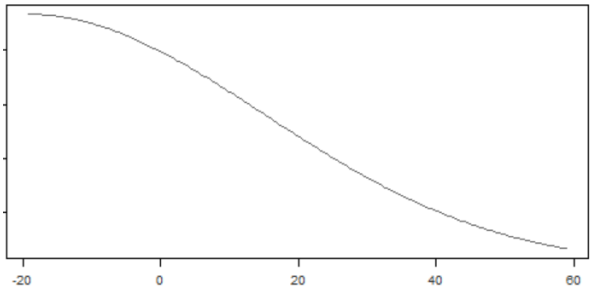


Figure 39. ROI Probability Distribution for Semi-Detached Score

3) *Terraced*

The terraced property provides the best performances as well as a decent number of cases. The median return based on test data is 12.89% with an accurate PROFIT classification of 95.47%. An impressive 1.94% of cases present returns exceeding -5% in LOSS while 84% exceed 5% in profit. 65% of cases surpass 9.71% in return and 63.55% yield a return greater than or equal to 10%.

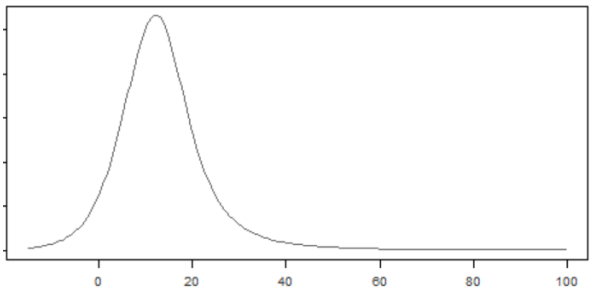


Figure 40. ROI Probability Distribution for Terraced Test

An almost symmetrical distribution is observed for returns on the score dataset which already hint at poor performance. Median return is at 1.52% with 65% of cases exceeding 0.80% and 5.61% below 5% in return. A huge drop in 5% and 10% returns with probabilities at 28.10% and 12.12% respectively.

The summary of the result can be characterized as one which did not provide a highly profitable return as opposed to poor classification as accuracy remained at a modest 75.76%.

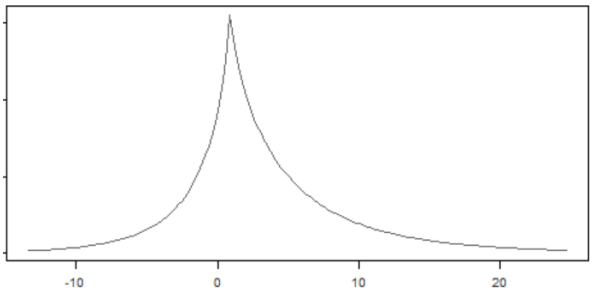


Figure 41. ROI Probability Distribution for Terraced Score

4) *Flat*

Results for the Flat model show extremely promising data. The test data provides median returns of 12.75% with 65% of cases surpassing 9.60% in return as well as an accurate PROFIT classification of 91.87%. Cases below -5% in return account for 3.63%, above 5% are 82.26% of the distribution and those exceeding 10% are 63%.

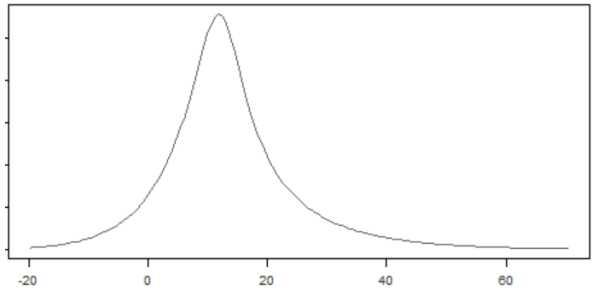


Figure 42. ROI Probability Distribution for Flat Test

Unfortunately, an attempt to score the Flat model failed with only 4 cases attaining a probability level for PROFIT over 85%.

5) *Test*

A summary of the test data for all models shows very promising features. A notable feature is skewness and right tail. As seen in figure 43, the majority of the distribution is within a range of -10% to 40% with the right tail highlight probable outcomes as high as 150%. This shows success in terms of modeling both the binary variable and its parent interval variable.

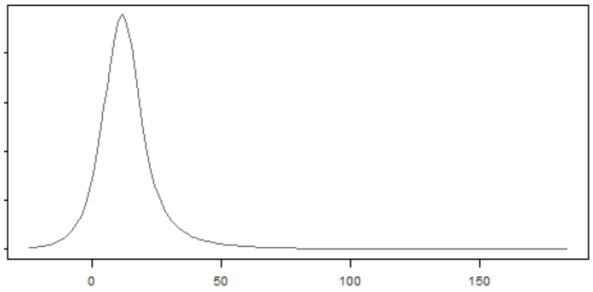


Figure 43. ROI Probability Distribution for Test

6) *Score*

High expectations for the model via test results were dampened by outcomes derived for the score dataset. An assessment of the poor performance would hint at issues with the year 2018 and market sentiment for investors heading into the year 2019. Only the terraced model performed decently in areas of risk management but remained poor relative to the high-profit yield test data.

The score distribution is normally distributed between [-20%, 20%] hence the main problem which needs to be scrutinized. The right tail property can be dismissed as the market maintains a 2:1 PROFIT to LOSS ratio hence a likelihood for higher return remains therefore, the median return of 4.02% is a point of emphasis.

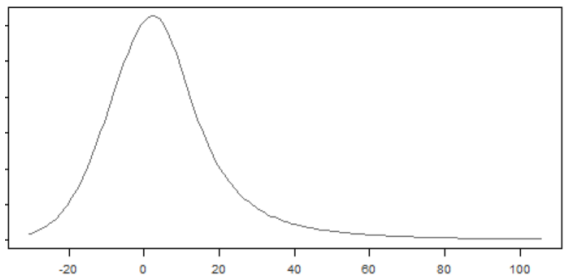


Figure 44. ROI Probability Distribution for Score

7) Model Performance

Evaluating the relationship between the probability level for PROFIT and corresponding Forecast (return) using centile fans produces an interesting trend. The centile fans is based on a model fitting the data according to the location (mean), scale (variance), and shape (skewness and kurtosis). The plot below demonstrates the probable outcomes of return relative to fitted probability hence a trend is realized with both variables being directly proportional in terms of fitted values. However, variance and shape reduce with an increase in gradient as higher returns are reserved strictly for probability levels for PROFIT above 96%.

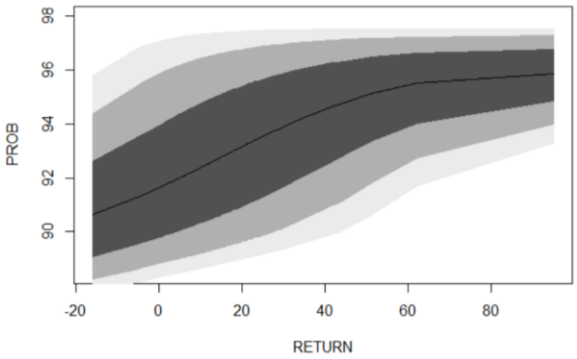


Figure 45. Centile Fan Illustrating Probability against Return

An almost uniform distribution observed for the test data for the probability level for PROFIT. This provides a rationale for higher median values in test data relative to those of the score dataset as probabilities above 94% are present.

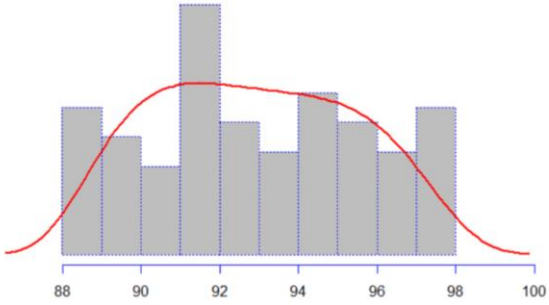


Figure 46. Probability Distribution for Test

Distribution of the score dataset's probability level for PROFIT provides a fitting rationale for the poor performance in 2019 by 2018 predictions. The majority of probability levels reside between [85%, 91%] which justifies the discrepancy between test and score performance.

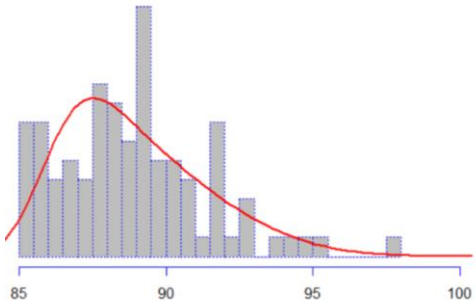


Figure 47. Probability Distribution for Score

TABLE 3

ACCURACY OF MODELS BASED ON TEST AND SCORE DATA

Accuracy				
Accuracy	Flat	Terraced	Semi Deta	Detached
Test	91.87%	95.47%	92.97%	76.47%
Score	50%	75.76%	69.23%	63.64%

TABLE 4

Probability of returns < -5%				
Accuracy	Flat	Terraced	Semi Deta	Detached
Test	3.63%	1.94%	3.08%	12.71%
Score		5.61%	32.31%	29.11%

TABLE 5

Probability of returns > 0.00%				
Accuracy	Flat	Terraced	Semi Deta	Detached
Test	91.85%	94.22%	91.93%	79.60%
Score		74.18%	57.35%	61.33%

TABLE 6

Probability of returns > 5.00%				
Accuracy	Flat	Terraced	Semi Deta	Detached
Test	82.26%	84.00%	80.55%	69.78%
Score		28.10%	47.82%	51.89%

TABLE 7

Probability of returns > 10.00%				
Accuracy	Flat	Terraced	Semi Deta	Detached
Test	63.00%	63.55%	60.75%	58.46%
Score		12.12%	39.23%	43.08%

8) Conclusion

The resulting lack of highly probable postcode sectors from the 2018 end of year data for 2019 predictions is ultimately down to the Neural Network model and its weightings. Due to the vast majority of score data missing highly probable outcomes, it stands to reason the cause will be a variable with whose segment either encompasses Boroughs or London as a whole. Variables within the model whose segments are classified under Boroughs include:

- DWELLD - Dwellings per Hectare
- WEARN – Workplace Earnings
- BVOR – Net Change in Borough Sales
- MTAX – Net Change in Median Council Tax

Credit is the only variable that incorporates the whole of London and is translated as MRGI which is the percentage change in year-on-year mortgage interest rate.

Translating the weighting on the Neural Network, the most significant hidden layer is the second (H12) which negates the output but does so with greater magnitude relative to H11. The variable which has the most effect on the hidden layer H12 is MRGI.

In 2018, the MRGI at the end of the year was a 6.7% rise in the mortgage interest rate from 3.58% to 3.82%. Although credit may not be the only reason behind the 2019 market performance, it does explain why highly probable data points were not predicted for the score dataset.

V. DEPLOYMENT

A. Deployment Strategy

The best manner through which this model can be implemented is by utilization by investors with either a vast amount of capital for procurement of properties in multiple locations and different types or by an investor with assess to credit to leverage her/his capital.

Diversification of assets acquired significantly reduces risk hence making it essential that numerous postcode sectors are selected for acquisition. Furthermore, the model should be used for assisting a domain property investment analyst and not relied on as the main criteria for investments. Numerous property types and locations are available for further filtering by a group/company who has experience in buying, selling, and leveraging Property in the London Housing Market.

The length expected for holding investments is a year which qualifies the model as a short-term investment strategy.

B. Monitoring and Maintenance

Although models were trained with year-on-year data, it is important to adapt quarterly data where benchmarks begin at the end of the previous year to Q1/Q2/Q3 as the probability of selected properties are constantly monitored to note variations as well as corresponding trends in house price changes.

Maintenance will involve the retraining of the model to refit variable weightings to adapt to the constant change in market cycles. This refitting would be done using the most recent 4-year market cycles meaning it will be done yearly with changes in model properties properly distinguished and accounted for.

VI. BENEFITS AND COMMERCIAL RISKS

A. Benefits

Proper implementation of results ought to benefit an investment group by supplementing discretional analysis with quantitative and statistical reasoning. Transformation of asset acquisition will be inevitable as critical factors for speculation are acknowledged hence delivering an edge over the market.

The acknowledgment of deciding factors could also assist in the discretional analysis as factors such as Transport Distance, Rent to House Price and House Price to Borough price can be easily understood to rationalize speculation. Furthermore, the acknowledgment of defining factors can ease transactions due to credible reasons for selling properties for profit to a potential buyer.

B. Commercial Risks

1) Data Quality

Firstly, binary classification is for estimating trends are not the direct return on a property. The average house price used is subject to its distribution as scrutiny based on the quality of the property meaning the % return signifies an increase in average price valuations and not the increase in all properties within a specific property in a postcode sector.

Secondly, Borough data limit sample sizes as seen in fig 17. of MTAX. Limiting certain variables to boroughs does not adequately narrow down a properties trend hence a possibility of clusters in probability may occur in cases where significant Borough changes affect the output classification.

2) Lack of Flexibility

The adaptability of a model to irregularities is not implemented in this case. Therefore, in cases where the market reacts to issues such as a pandemic or recession, the model may find it difficult to quantify irregularities in market patterns.

3) Risk-to-Reward

A ratio of the risk to reward of an investment will highlight the return on every loss. This will be done by accepting a benchmark for loss and profit.

The benchmark for a loss will be -5% which means once a property loses 5% of its valuation at the point of acquisition, it will be sold to limit risk. A similar process occurs with profit, a property is sold once it gains 10% of its valuation. For prices which do not meet benchmarks values, the following will be assumed:

- -2.50% return for [-5.00%, 0.00%]
- 2.50% return for [0.00%, 5.00%]
- 7.50% return for [5%,5.00%]

The assumption to be made will be a uniform distribution of capital over 100 properties with a total allocation of £100,000,000.

As seen in Table 8 test data provides an estimated return of 7.56% with a total loss of £306,750 and a profit of £7,865,250 making the risk-to-reward ration approximately 26/1. Considering an accurate PROFIT classification of 91.48% for test data, the risk-to-reward ratio is justifiable.

TABLE 8

ESTIMATED ROI BASED ON TEST DATA DISTRIBUTION

Test Probability					
Accuracy	< -5.00%	[-5%,0%]	[0%,5%]	[5%,10%]	> 10%
Return	-5%	-2.50%	2.50%	7.50%	10%
Test	3.75%	4.77%	10.66%	19.33%	61.49%
Test Capital	£3,750,000.00	£4,770,000.00	£10,660,000.00	£19,330,000.00	£61,490,000.00
Test Return	-£187,500.00	-£119,250.00	£266,500.00	£1,449,750.00	£6,149,000.00

The table below shows return estimates with the score data based on the assumption stated. A return of 3.87% is calculated with a loss at £1,232,250 and profit at £5,099,500 meaning a risk-to-reward ratio of 4/1 is calculated with a model accuracy of 69.17%.

TABLE 9

ESTIMATED ROI BASED ON SCORE DATA DISTRIBUTION

Score Probability					
Accuracy	< -5.00%	[-5%,0%]	[0%,5%]	[5%,10%]	> 10%
Return	-5%	-2.50%	2.50%	7.50%	10%
Score	18.46%	12.37%	19.33%	14.71%	35.13%
Score Capital	£18,460,000.00	£12,370,000.00	£19,330,000.00	£14,710,000.00	£35,130,000.00
Score Return	-£923,000.00	-£309,250.00	£483,250.00	£1,103,250.00	£3,513,000.00

REFERENCES

[1] J. Marsden, "House prices in London – an economic analysis of London's housing market", 2015. [Online]. Available: https://www.london.gov.uk/sites/default/files/house-prices-in-london.pdf. [Accessed: 11- Jun- 2020].

[2] HM Land Registry, "HM Land Registry house price data", Landregistry.data.gov.uk, 2020. [Dataset]. Available: https://landregistry.data.gov.uk/app/standard-reports. [Accessed: 10- Jun- 2020].

[3] Bank of England, "Standard Variable Mortgage Rates to Households", Bankofengland.co.uk, 2020. [Dataset]. Available: https://www.bankofengland.co.uk/boeapps/database/fromshowcolumns.asp?Travel=NIxAZxSUx&FromSeries=1&ToSeries=50&DAT=RNG&FD=1&FM=Jan&FY=2010&TD=11&TM=May&TY=2025&FNY=Y&CSVF=TT&html.x=66&html.y=26&SeriesCodes=CFMBX2D&UsingCodes=Y&Filter=N&title=CFMBX2D&VPD=Y. [Accessed: 10- Jun- 2020].

[3] Bank of England "Standard Variable Mortgage Rates to Individuals and Individual Trusts", Bankofengland.co.uk, 2020. [Dataset]. Available: https://www.bankofengland.co.uk/boeapps/database/fromshowcolumns.asp?Travel=NIxAZxSUx&FromSeries=1&ToSeries=50&DAT=RNG&FD=1&FM=Jan&FY=2010&TD=11&TM=May&TY=2025&FNY=Y&CSVF=TT&html.x=66&html.y=26&SeriesCodes=CFMBX2D&UsingCodes=Y&Filter=N&title=CFMBX2D&VPD=Y. [Accessed: 10- Jun- 2020].

[5] Greater London Authority "Land Area and Population Density, Ward and Borough – London Datastore", Data.london.gov.uk, 2020. [Dataset]. Available: https://data.london.gov.uk/dataset/land-area-and-population-density-ward-and-borough. [Accessed: 11- Jun- 2020].

[6] UK Finance "Mortgage lending within UK postcodes | UK Finance", Ukfinance.org.uk, 2020. [Dataset]. Available: https://www.ukfinance.org.uk/data-and-research/data/mortgages/mortgage-lending-within-uk-postcodes. [Accessed: 11- Jun- 2020].

[7] Transport for London "Public Transport Accessibility Levels – London Datastore", Data.london.gov.uk, 2020. [Dataset]. Available: https://data.london.gov.uk/dataset/public-transport-accessibility-levels. [Accessed: 11- Jun- 2020].

[8] Greater London Authority "Total Number of Dwellings and Net Additional Dwellings, Borough – London Datastore", Data.london.gov.uk, 2020. [Dataset]. Available: https://data.london.gov.uk/dataset/net-additional-dwellings-borough. [Accessed: 11- Jun- 2020].

[9] Office of National Statistics" Earnings by Workplace, Borough – London Datastore", Data.london.gov.uk, 2020. [Dataset]. Available:.https://data.london.gov.uk/dataset/earnings-workplace-borough. [Accessed: 11- Jun- 2020].

[10] Office of National Statistics "Earnings by Place of Residence, Borough – London Datastore", Data.london.gov.uk, 2020. [Dataset]. Available: https://data.london.gov.uk/dataset/earnings-place-residence-borough. [Accessed: 11- Jun- 2020].

[11] Office of National Statistics "Employment Rates by Gender and Age or Disability, Borough – London Datastore", Data.london.gov.uk, 2020. [Dataset]. Available: https://data.london.gov.uk/dataset/employment-rates-by-gender-and-age. [Accessed: 11- Jun- 2020].

[12] Ministry of Housing, Communities and Local Governments "Council Tax Charges – Bands, Borough – London Datastore", Data.london.gov.uk, 2020. [Dataset]. Available: https://data.london.gov.uk/dataset/council-tax-charges-bands-borough. [Accessed: 11- Jun- 2020].

[13] Valuation Office Agency "Average Private Rents, Borough – London Datastore", Data.london.gov.uk, 2020. [Dataset]. Available:

https://data.london.gov.uk/dataset/average-private-rents-borough. [Accessed: 11- Jun- 2020].

[14] G. Rehkemper, London Wards and Boroughs 2015. 2015

[15] Office of National Statistics, London Boroughs (December 2017) Map in London. 2017.

[16] J. Quinlan, C4.5. San Mateo, Calif.: Morgan Kaufmann Publishers, 1993.

[17] M. Stasinopoulos, *Flexible regression and smoothing*, 1st ed. Chapman and Hall/CRC, 2017.

.