Evaluation Metrics based comparative analysis of machine learning algorithms

*

Davis Munachimso Agughalam
School of computing
National college of Ireland
Dublin, Ireland
agughalamdavis@yahoo.com

***Abstract*—The use of machine learning techniques have cut across various domains especially with increase in amount of data available and improvements in computing. There is no one algorithm that is better than the rest in machine learning as performance is highly case dependent and most times depends on the interpretation. This gives rise to the need to use more than one algorithm and compare based on evaluation metrics. This research carries out a comparative analysis between machine learning algorithms applied across different domains, housing, ecommerce and credit default. The work measures the performance of these algorithms on evaluation metrics such as mean squared error, accuracy, kappa, sensitivity and specificity as interpreted for their areas of application.**

***Index Terms*—Machine learning, accuracy, sensitivity, specificity, credit default, ecommerce, housing**

I. Introduction

Backed by the increase in availability of data and computing power, leveraging machine learning techniques for the discovery of patterns in data has enabled a new wave of research in different sectors for both predictive and classification analytics. The use of machine learning algorithms have currently cut across so many areas of application leading to the use of several algorithms. Selection of an algorithm to use for a particular scenario has become a little difficult due to the diversity and suitability of these algorithms for different purposes. Some algorithms are more suited to particular areas of applications than others thus creating the need for a comparative analysis of these algorithms. This process is evident across various machine learning researches as different models are used to solve particular problems and compared against based on standardized evaluation metrics. In order to cover a wider scope of application, three different sectors, housing, banking and ecommerce are cut across in this research. For housing, application of machine learning algorithms to predict house rental prices is the focus. This is an important area of research especially with increasing urbanization and consequent growing demand for housing. Real estate developers can leverage the models to generate a satisfactory pricing strategy while buyers can also use the model for insights. In the banking sector, one of the most important areas of operation is the loan system through credit cards. However, this is abound with risk as there is a chance customers can default on these loans leading to heavy losses. This is one of the many areas machine learning is currently being applied to help determine the likelihood of a client defaulting or not defaulting on the credit card. This provides a perfect opportunity to compare machine learning algorithms. Another area with huge potential for use of machine learning is in ecommerce. Many ecommerce stores have used machine learning to be able to improve customer service and consequently drive higher profit margins. Being able to ascertain the probability of making revenue from a potential customer visiting the ecommerce platform would enable more customized and targeted advertising thereby ensuring more profits. Machine learning algorithms can also be leveraged for customer segmentation to enable targeted advertisements and promotions to also increase profitability. By investigating the use of machine learning in different domains, this research is mainly themed around comparing the performance of these algorithms. It tries to expose the strengths and weaknesses of selected machine learning algorithms by exhaustively analyzing and comparing them in the domain of application based on already standardized evaluation metrics. The the paper is structured as follows: Section II gives the background information on the algorithms used in the research. In Section III, the related work is exhaustively and critically analyzed. The methodology and results are explained in Section IV while conclusion in section V.

II. Background

This section outlines important background information about the machine learning algorithms applied for this research.

A. Machine Learning Algorithms

Machine learning involves building models that learn patterns from past data to make future predictions or gain insights about the past [1]. These algorithms are broadly divided into two categories including:

- **Supervised Learning:** Supervised learning algorithms are machine learning algorithms applied when there is an outcome variable in the data that can be predicted or classified by other independent variables in the data set[1]. These are broadly divided into two categories, namely classification and regression. Regression algorithms are employed when the target variable is continuous, numeric data while the classification algorithms are used when the target variable is categorical, and discrete. The supervised algorithms used in this research are decision tree regression, linear regression and random forest for regression analysis while decision trees, K-nearest Neighbours (KNN), Support Vector Machines (SVM) and Logistic regression are used for classification.

- **Unsupervised Learning:** Unsupervised learning algorithms are applied when there is no target or outcome variable. They are mainly used to gain insights or make inferences rather than outright predictions or classification and are some times applied to label the data for supervised learning algorithms[1]. The algorithms group the data into clusters based on similarities within the data values.

B. Regression Algorithms

Regression algorithms are used for predictions when the outcome variable is numeric or quantitative data. For the purpose of prediction of house prices, the data to be predicted is numeric thus suitable for applying regression algorithms. The regression algorithms used for the prediction are;

- **Multiple Linear Regression:** Linear regression models carry out prediction by creating a linear model for the data. This model identifies the functional relationship between the predicted variable and the independent variables[1]. In other to ascertain the best possible relationship between data points, it leverages the ordinary least squares estimation technique to fit a line through the data. The equation for the fit line is the relationship between the outcome and the predictor variables. It is mathematically expressed as;

  $$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{1}$$

  where y is the dependent variable, X_1, X_2 and X_n are predictor variables, β_0 the intercept, and β_1, β_2, β_n are coefficients.

- **Decision Tree Regression:** The decision tree can be viewed as a continuous set of yes or no questions that are used to achieve the final aim of prediction [1]. A question is determined at each node that leads to the reduction of the gini impurity when answered. The decision tree iterates through feature values on which splits are made to ensure as low a gini impurity value as possible. The Gini Impurity of a node can be defined as the probability that a random chosen sample in a node would have an incorrect label if sample distribution in the node was used to label [1]. As the divides downwards, the gini impurity reduces till the final predictions.

- **Random Forest Regression:** The random forest model is an aggregation of decision trees with added functionality for improved performance [2]. It randomly samples the data for building each decision tree thereby building each tree with a slightly different set of data than the others. A subset of features is selected randomly for splitting at a node. Drawing the samples with replacement gives rise to the term bootstrap aggregation, also known as bagging [2]. The final decisions on the random forest are then made by averaging the predictions of all contributing decision trees.

These three algorithms in their base and advanced forms through parameter tweaking are employed for prediction of house rental prices and compared against each other in terms of the root mean squared error (RMSE), mean squared error (MSE) and mean absolute error (MAE).

C. Classification Algorithms

Classification algorithms are used to predict discrete qualitative data. They are mostly used to predict in which group or class a particular data point falls in after learning from historical data. The classification algorithms for this research are as follows;

- **Logistic Regression:** Logistic regression algorithms are classification algorithms used to predict the probability of a data entry belonging to a particular class [2]. Logistic regression models can be binary (2 classes) or multinomial (More than 2 classes). Just like the linear regression model, they create relationships between the predicted variable and the predictor variables and use this relationship to determine the probability of belonging to a particular class. It can be mathematically expressed as;

  $$p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \tag{2}$$

  where p equals probability of belonging to the class, β_0 equals intercept, β_1 is the coefficient of the independent variable and X is the independent variable.

- **Support Vector Machine (SVM):** An SVM can be conceptualized as a boundary creating surface between two points of data plotted in a multi dimensional space representing feature value examples [1]. It can be used for both classification and regression problems. It does classification by creating a hyperplane to divide the input data into fairly homogenous partitions or classes depending on the similarity values within the features. SVMs can be employed for both linearly seperable and non linearly separable data through the use of kernels. The hyperplane is chosen by taking into consideration the distance between the closest data points on either side. This distance is called the margin and the data points are known as support vectors. This process is iterative until maximum distance between the points and the hyperplane is achieved [1]. The hyperplane at this stage is the optimal hyperplane. This model is very suited for application to the datasets for classification due to its ability to correctly classify large data.

- **K-Nearest Neighbours (KNN):** The KNN classification algorithm uses the distance between data points for classification. Data points that are close to each other are adjudged to be in the same class. This distance can be euclidean distances or any other notion of distance preset. It classifies a data point into a class by taking into consideration the distance between the data point and k number of nearest data points in the plane. There is no model produced at the end of the classification hence it is known as a lazy learner and this lack of an output model at the end of the classification process also takes away from the interpretability of the results beyond classification [1].

Two of the classification algorithms, support vector machines and logistic regression are used and compared on the ecommerce dataset while the K-nearest neighbours algorithm and the classification variant of the decision tree algorithm are used for the credit default analysis.

D. Evaluation Metrics

To evaluate algorithm performance on the prediction or classification purpose, certain metrics are used. The metrics for evaluating classification and prediction algorithms are different. For the regression part of this research, the following metrics are used for evaluation.

- **Mean Squared Error (MSE):** The MSE measures the average of the differences between the actual value and the predicted values of the data points [1]. Usually, the lower the MSE, the better the model as it shows that difference between predicted and actual values is low. Mathematically, it is expressed as;

$$MSE = \frac{1}{N}\sum (y_i - \hat{y_i})^2 \tag{3}$$

- **Mean Absolute Error (MAE):** The MAE is the average of the absolute differences between the predicted outcome and the actual outcome. The MAE is more sensitive to outliers than the MSE [1].

$$MAE = \frac{1}{N}\sum |y_i - \hat{y_i}| \tag{4}$$

- **Root mean Squared Error (RMSE):** The RMSE is derived by taking square root of the MSE. it represents average distances between the actual and predicted values in their actual units of measurements as the square used to remove the signs in the MSE is reversed with the square root. it is expressed mathematically as;

$$RMSE = \sqrt{\frac{1}{N}\sum (y_i - \hat{y_i})^2} \tag{5}$$

For the classification algorithms, some of the metrics used for evaluation are derived from the confusion matrix. The confusion matrix is a diagnostic tool for gaining insights into the kind of errors being made by the model[1].

TABLE I
CONFUSION MATRIX

	Predicted(1)	Predicted(0)
Actual(1)	True Positive	False Negative
Actual(0)	False Positive	True Negative

- **Accuracy:** This is the ratio of the number of correct predictions to the total number of predictions. For highly imbalanced data, this evaluation metric alone is not enough to determine model performance. This is because guessing the class with a higher count will still yield a high accuracy [1].

$$Accuracy = \frac{Number of correct predictions}{Total number of predictions} \tag{6}$$

- **Kappa:** The kappa statistic adjust accuracy by taking into account the role of chance in a correct prediction [1]. It is very important for data with high class imbalance as accuracy is not good enough in that scenario. Higher kappa values generally indicate better models.

$$Kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{7}$$

- **Precision:** The precision of a classification algorithm is the ratio of the true positively labelled predictions to all the positively labelled predictions . It is the proportion of positives that are really positive [1]. It tends towards minimizing false positives.

$$Precision = \frac{Truepositives}{Truepositives + FalsePositives} \tag{8}$$

- **Specificity:** The specificity of a model measures the amount of negative data points that were correctly classified [1]. It is the true negatives to the sum of the true negatives and the false positives ratio.

$$Specificity = \frac{Truenegatives}{Truenegatives + FalsePositives} \tag{9}$$

- **Sensitivity:** The sensitivity of a model measures the amount of positive data points that were correctly classified [1].

$$Sensitivity = \frac{Truepositives}{Truepositives + Falsenegatives} \tag{10}$$

- **F-measure:** The F-measure is obtained by calculating the harmonic mean of the precision and recall.

$$F - measure = \frac{2 * precision * recall}{recall + precision} \quad (11)$$

- **Receiver Operating Characteristic (ROC) curve/Area under the Curve (AUC):** The ROC curve is shown on a plot of the true positive rate against the false positive rate calculated at each iteration of the classification threshold. The classification threshold starts out at 50 percent and is steadily increased during this process and the area under the curve can be calculated. Higher AUCs indicate better models.

III. Related Work

Related work on application of machine algorithms and consequent evaluation of these algorithms are presented comprehensively in this section. Many researchers have looked into the use of machine learning algorithms for credit default analysis as this is an important area. One of such is the work done by Sayjadah et al.[3]. They applied 3 machine learning algorithms logistic regression, random forests and decision trees for the prediction of credit default with further evaluation and comparison of these algorithms using accuracy and area under the curve (AUC). All 3 algorithms had relatively close accuracy scores with the decision tree algorithm just edging it slightly. This was not the case for the AUC as the logistic regression algorithm had a higher AUC than the others hence was stated to be a better model in their analysis. This research used these algorithms in their base forms without any form of model improvements. Birla et al. [4] researched the impact of imbalanced data on credit risk prediction. They employed logistic regression and classification and regression trees (CART) coupled with techniques to deal with imbalanced data such as undersampling, prior probablities and matrix weighing. Of the methods to deal with the imbalanced data, the prior probablities approach caused the highest accuracy in the CART algorithms. The cloglog method of logistic regression also produced higher accuracy and AUC values. Their research buttressed the impact of data imbalance on the performance of machine learning algorithms but was limited to only two algorithms. Turkson et al. [5] also carried out research on using machine learning to predict bank credit worthiness by applying 7 ensemble algorithms namely extra tree, Random Forest, Random Forest (with only 5 features), AdaBoost, Gradient Boosting, Random Trees Embedding, voting and 9 other algorithms, Logistic regression, SVM, classification and regression tree (CART), Nearest centroid, KNN, Gaussian Naïve Bayes, Neural networks, linear discriminant analysis and bagging. These algorithms were evaluated and compared using accuracy, precision, recall, specificity and F-measure (F1). The algorithms performed reasonably well in terms of accuracy having an accuracy between 76 to 80 percent except for the nearest centroid and Gaussian naïve bayes. For the other evaluation metrics, the ensemble algorithms had relatively higher scores thereby showing that they are better algorithms for predicting credit worthiness. Credit risk analysis in peer to peer lending system was the focus of the research done by Kumar et al. [6]. Decision tree, random forest and bagging (bootstrap aggregation) algorithms were used and evaluated based on precision, accuracy, Receiver operating characteristic (ROC) and Area under the curve (AUC). The decision tree algorithm had highest precision values while the random forest and bagging algorithms had the same accuracy values higher than the decision tree. The decision tree also had higher AUC scores than the other algorithms hence doing a better job of classifying credit scores. In the aspect of housing, machine learning techniques are being applied for prediction of exact prices of the houses as well as the direction of the house price, that is, whether the prices are likely to increase or decrease. Banerjee and Dutta[7] researched the prediction of the direction of house prices using machine learning techniques. SVM, random forests and artificial neural networks were used for the analysis and evaluated on accuracy, precision, specificity and sensitivity. The random forest algorithm had the highest accuracy but was stated to be prone to overfitting, hence the SVM algorithm with a sensitivity value of 84 percent was a better algorithm compared to the other two. Feng et al. [8] explored application of machine learning algorithms to predict the compressive strength of concrete using AdaBoost comparing the results with SVM and Artificial neural network (ANN). This comparison was done based on metrics R2, root mean squared error (RMSE), mean absolute percentage error (MAPE) and mean absolute error (MAE). The AdaBoost algorithm was found to perform better than the other algorithms and this was attributed to the fact that it was an ensemble algorithm while others were single algorithms. Further comparison was done with previous studies where Logistic regression and CART were used and the AdaBoost algorithm still had a higher R2. Machine learning algorithms have also been used for health research. Kim et al. [9] carried out research for detection of acute hemorrhage shock in rats. They applied algorithms Artificial neural networks (ANN), logistic regression (LR), support vector machine (SVM), and random forest (RF). The experiments were carried out on 36 rats under anesthesia grouped in 3 depending on the blood loss volume. Heart rates and blood lactate concentration measurements amongst others were taken to serve model predictors. In terms of sensitivity and AUC values, The SVM model performed better. The research was limited because of the anesthetics to stabilise the subjects. The experiment recordings might be different under normal circumstances due to the impact of the drugs on the cardiovascular system. Isuflorane anesthesia was suggested for future work as this could mimic normal conditions and provide stable results. Santos et al. [10] looked into using machine learning for prediction of the 30 day adjusted survival in critically ill cancer patients in a study carried out in Brazil. Penalized logistic regression, Logistic regression, gradient boosted trees, basic decision trees, artificial neural networks and random forest were used and compared using the AUC/ROC curve. All the algorithms except basic decision trees offered high predictive performances for the scenario with AUC/ROC values of over 0.8 with the ANN and gradient boosted trees having the best overall calibration. Lopes et al. [11] used machine learning to automate orthogonal defect classification in computer software systems. They split the data into subsets of attributes namely activity, code inspection, function test, system test, unit test, impact and target, defect type, and qualifier. They applied KNN, SVM, random forest, Naïve Bayes, nearest centroid and recurrent neural networks and compared them using accuracy, recall and precision for each of the subsets. The SVM algorithm showed highest scores of precision, accuracy and recall across all the data subsets thereby doing a better job of classification. Ma et al. [12] used machine learning for prediction of risk of acute renal failure. The algorithms applied were and compared were, logistic regression, C5.0 decision tree, and artificial neural networks. The decision tree performed best to predict death and dialysis scenarios. Their results further suggested that the decision tree model can be used for real life diagnosis as it is interpretable. Entekhabi et al. [13] explored machine learning as a medium of predicting the rate of degradation of genipin cross-linked gelatin scaffolds by using kernel ridge regression and artificial neural networks (ANN) algorithms. These algorithms were validated and compared using MSE as an evaluation metric and it was found that the ANN algorithm performed better with an MSE of 2.68 percent compared to 4.78 percent MSE for the kernel ridge regression. Babu and Beulah [14] in their research used machine learning algorithms to predict air quality. Logistic regression, random forest, decision tree, k-nearest neighbors (KNN), and support vector machine methods were employed and compared using the accuracy, precision, recall and F1 score. They concluded that the decision tree algorithm worked best from the results gotten from all four evaluation metrics. Chen et al. [15] further explored harnessing the predictive capabilities of machine learning algorithms in the domain of cryptocurrency. In their work to predict bitcoin prices, they employed methods such as logistic regression, linear discriminant analysis, random forest, XGBoost, quadratic discriminant analysis and support vector machines (SVM) and compared these methods in terms of the accuracy and precision. The SVM algorithm was shown to have highest accuracy and precision values. Integration of machine learning techniques and physiology based heart rate features for antepartum fetal monitoring was the focus off the research by Signorini et al. [16]. They leveraged 15 machine learning algorithms for this purpose and carried out a comparison using accuracy. Random forest performed better than the other models with an accuracy of 0.911 even though classification trees, logistic regression and SVM had close values. Research aimed at the use of machine learning models for prediction of seasonal rainfall extremes of peninsular Malaysia was done by Pour et al [17]. SVM, RF, and Bayesian artificial neural networks (BANN) were used and it was found that the BANN outperformed other algorithms across all evaluation indices used. No matter the domain of application, the central theme of employing more than one algorithm and comparing results based on evaluating metrics as evident in the above researches goes to show that the superiority of one machine learning model over another is highly case dependent as different data sets have different characteristics that can impact a machine learning model positively or negatively going by the evaluation metrics. This research hopes to further buttress that point by building and evaluating a host of machine learning models across different domains.

IV. Methodology

A. Data Mining Process

In order to answer the research question systematically, the Knowledge discovery in databases (KDD) process was followed. This process is elaborated in the following steps;

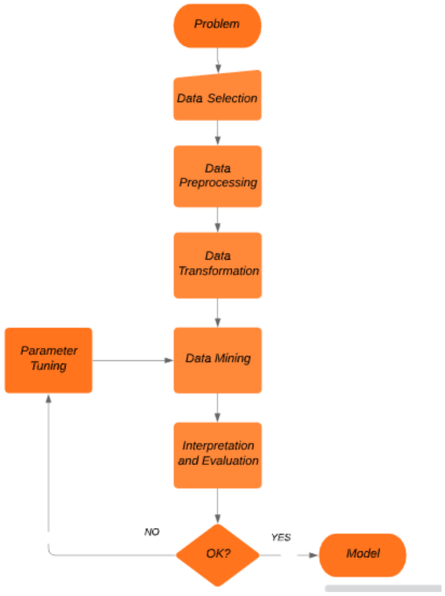

Fig. 1. Machine learning process flow

- **Selection:**This step involves selecting the required dataset and highlighting the features or a subset of variables required from the dataset for the analysis and subsequent knowledge discovery.
- **Preprocessing:**Data cleaning and preprocessing involves getting rid of noise and outliers and developing strategies for handling missing data to get consistent data.
- **Transformation:**This step involves transforming the data to suit required needs through dimensionality reduction, feature engineering or tackling data distribution through other transformation processes
- **Data Mining:**Choosing the data mining task (classification, regression or clustering) and the accompanying algorithm to search for patterns of interest is the aim of the data mining phase of the KDD process.
- **Interpretation and evaluation:** Interpretation of the mined patterns as relates the domain of application and further evaluation is the final step of the process.

As this research cuts across three different areas of application with different data requirements, the process enumerated above is applied for each of them.

B. House rental prices

1) Data Selection: The dataset used was gotten from Kaggle data repository, submitted by datafiniti [18]. The dataset contained 11,375 rows and 25 columns. Some columns contained metadata and other data irrelevant to the focus of the research and as such were removed from the dataset. The target dataset consisted of 8 columns after selection.

2) Data Preprocessing: Part of columns in the target dataset were renamed for ease of use, other columns such as the price column had both numeric data for the actual price and character data for the price unit in each entry. These were stripped out and only the numeric part of that data was left for analysis. The data columns were then converted to the proper data format as required for the analysis. Further data exploration was done on the dataset to search for missing values and take care of them either through imputing, removal or replacing with mean as the case requires. Fig 2 shows the data in each of the columns in the dataset. The numBed column had more than 50 percent of the data missing and as such was dropped. For the minStay and prices.price column, the missing data was minimal so the rows with the missing data were dropped in its entirety as it would have negligible impact on the analysis.

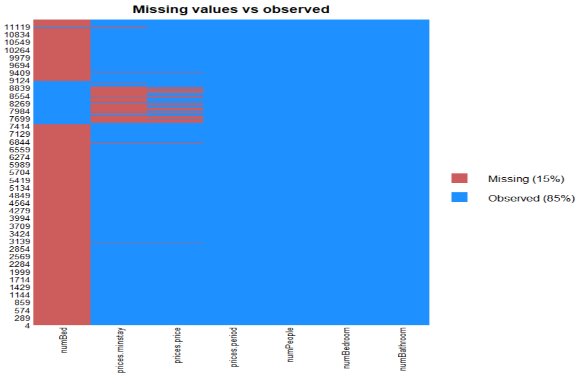

Fig. 2. Missmap showing missing data in each column

3) Data Exploration and Transformation: Predicting the rental price of the home depending on the other features was the objective of this data mining process. This meant that the prices.price column is the dependent variable and it would be important to see the distribution of the data. Making a histogram plot of the data showed that it was highly right skewed and as such a transformation was required. A log transform was performed to improve the distribution of the data.
The correlation of the predictor variables to the predicted variable was also checked and they were correlated enough for a valid regression analysis.

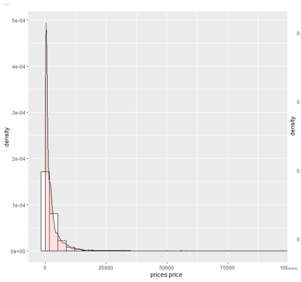

Fig. 3. Distribution of prices column

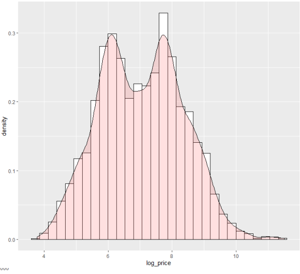

Fig. 4. Log transformed distribution of prices column

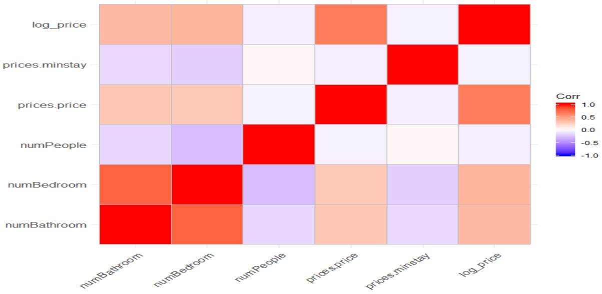

Fig. 5. Correlation plot

4) Data Mining: To perform the prediction task, linear regression, decision tree regression and random forests algorithms are employed and subsequently evaluated. A visualization of the decision making process of the decision tree is shown in fig 6 and the evaluation results of the 3 models are tabulated in table II.

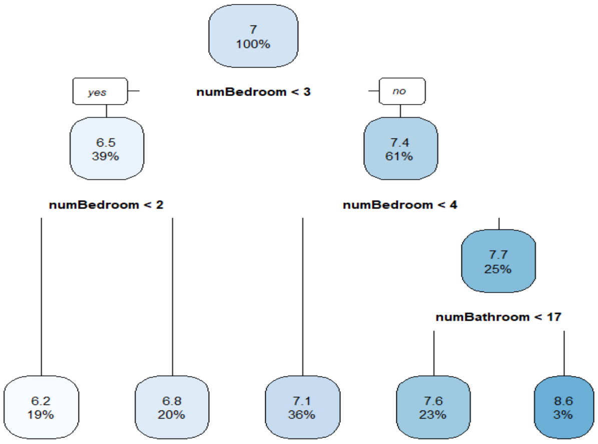

Fig. 6. Decision tree

5) Interpretation and Evaluation: The linear regression model had an R squared of 0.8545 showing that the predictors account for 85 percent of the variance within the independent variable. To check the performance of on the test data, the model was used to predict the outcome variable. The mean squared error was found to be 0.23 and the mean absolute

TABLE II
EVALUATION RESULTS

Model	MSE	RMSE	MAE
Linear regression	0.23	0.48	0.37
Decision tree	1.39	1.17	1.02
Random forest (bagging)	0.17	0.41	0.31

error, 0.37 and a root mean squared error of 0.4796. This showed that the model predicted values were close to the actual values but there was still room for improvement. The decision tree model had an MSE, MAE and RMSE of 1.39, 1.02 and 1.17 respectively. The predicted values were farther from the actual values than that of the predictions using the linear regression model. To see if another model could do a better job than both the linear and decision tree models, a bagged random forest model was built. The model had an MSE of 0.17, an MAE of 0.31 and an RMSE of 0.41 essentially outperforming the other two models in predicted values closer to the actual values. Amongst all 3 models, the bagged random forest algorithm did the best job of predicting the house rental prices.

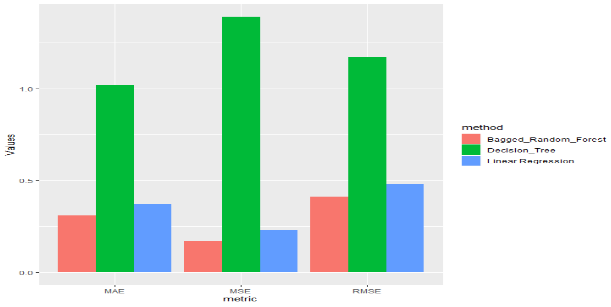

Fig. 7. Evaluation metrics comparison bar chart

C. *Credit card default*

1) Data Selection: The data for this analysis was gotten from the University of California Irvine (UCI) machine learning repository submitted by Yeh and Lien [19] in Microsoft excel format. The dataset contained information to determine whether a client was going to default on the credit card for the next month. The dataset had 30,000 rows and 25 columns. The “default payment next month” is considered the response variable in this analysis. All the other features of the dataset were relevant to the analysis and as such, no columns were removed.

2) Data Preprocessing: There were no missing data values. Further exploration showed that the response variable was highly imbalanced with about 77.88 percent saying no default while only 22.12 percent for the default. This imbalance would need to be taken care of during data mining. Feature selection was also done to select the best predictors to be used for the analysis.

3) Data Exploration and Transformation: The data set was further explored to find some more information. Fig 9 shows the distribution defaulters across different age groups and their balances (LIMITBAL). It is seen that people with lower balances are more likely to default across the age groups. Fig 10 shows the distribution of defaulters across different marital status. Singles and divorcees have the highest number of defaulters across all the marital groups considered.

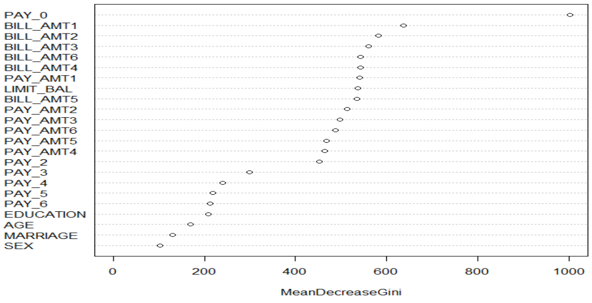

Fig. 8. Variable Importance

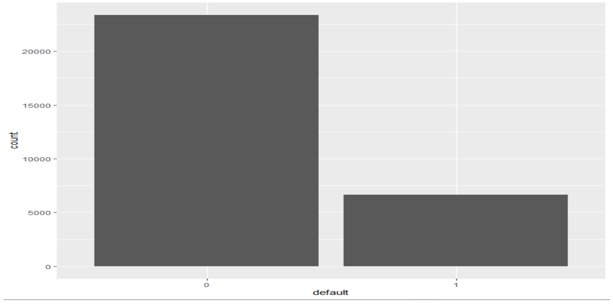

Fig. 9. Class Imbalance

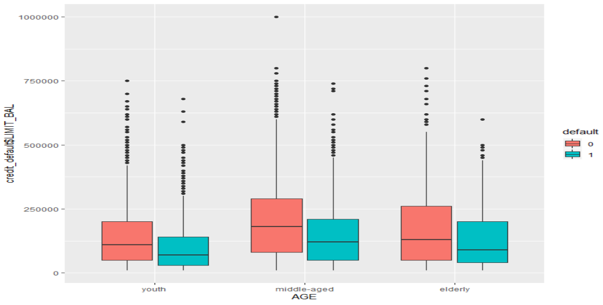

Fig. 10. Boxplots showing distribution of AGE, LIMITBAL and Default

4) Data Mining: To carry out the classification, the k-nearest neighbors algorithm and decision tree classification are used and their results compared to each other. For the KNN algorithm, the sub selected features from the preprocessing phase were used. To deal with the class imbalance and avoid model bias towards a class, the data used to train the model was upsampled and downsampled and their results compared. The caret package was used for parameter tuning for the k value, used to set up 10 fold cross validation and subsequent normalization of the data for analysis with the downsampled data. The model selected an optimum k value of 247 with the downsampled training data. The resulting model accuracy

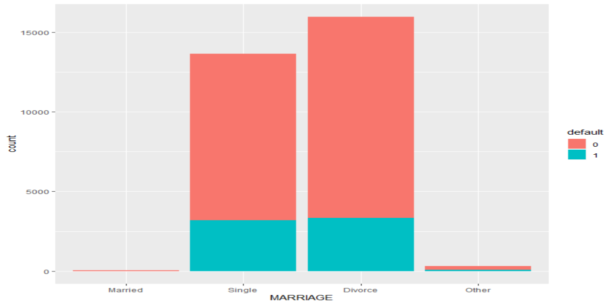

Fig. 11. Bar graph showing distribution of defaulters across marital status

was 72.57 percent having a kappa value of 0.3192. For the upsampled data, the caret package was not used for automated parameter tuning due to computational costs. Instead, a rule of thumb was used to test different values of K and an optimum value of 159 was found to produce an accuracy of 78.65 percent and a kappa value of 0.37 showing that upsampling was a better option than downsampling due to possible information loss. The decision tree algorithm produced an accuracy of 77.34 percent and 35.68 kappa value. The first tree used only 3 tree nodes and this was found to be the optimum number of tree nodes after pruning the tree.

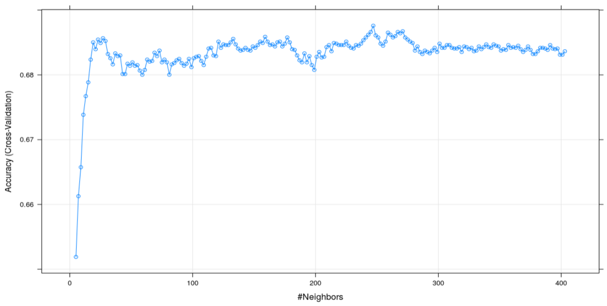

Fig. 12. K values for down sampled training data

5) Interpretation and Evaluation: The results from evaluation of the models are tabulated as follows.

TABLE III
KNN WITH UPSAMPLED DATA CONFUSION MATRIX

	Predicted(1)	Predicted(0)
Actual(1)	936	824
Actual(0)	1106	5806

TABLE IV
DECISION TREE CONFUSION MATRIX

	Predicted(1)	Predicted(0)
Actual(1)	985	985
Actual(0)	957	5645

The credit default research is mostly interested at identifying credit defaulters hence 1 is the positive class while 0 is the negative class. Wrongly classifying non defaulters would be less expensive than wrongly classifying defaulters. This means that the best model should be able to reduce false negatives as much as possible. Owing to this, sensitivity becomes an important metric in evaluating the model performance. The decision tree model, even though it had the lesser accuracy, was able to minimize the false negatives more than the KNN models. This makes it the better model in this instance as it would make lesser mistakes identifying defaulters as non defaulters. The kappa values of 0.3568 and 0.37 for both models also show that their predictions are in fair agreement with the actual values hence the models are satisfactory even though performance can be improved.

TABLE V
MODEL COMPARISON

Model	Accuracy	Kappa	Sensitivity	Specificity
KNN	78.65	0.37	0.481	0.8757
Decision tree	77.34	0.3568	0.5072	0.8514

D. Ecommerce

1) Data Selection: The dataset for the analysis was gotten from the UCI machine learning repository, submitted by Sakar et al. [20]. The dataset contained information for an ecommerce website's visitors and whether the visits were converted to revenue or not. It consisted of 12,331 columns and 18 rows with the response variable being "Revenue". All columns of the dataset were relevant to the research and as such, none was dropped.

2) Data Preprocessing: The dataset had no missing data. The response variable was imbalanced towards the no revenue class and this was taken into account and corrected during the data mining phase. As the SVM and logistic regression algorithms required categorical data to be encoded, the categorical data in the data columns were encoded through one hot encoding. This created a total of 77 features. Feature selection was done to select the best 12 predictors and these were used for both algorithms.

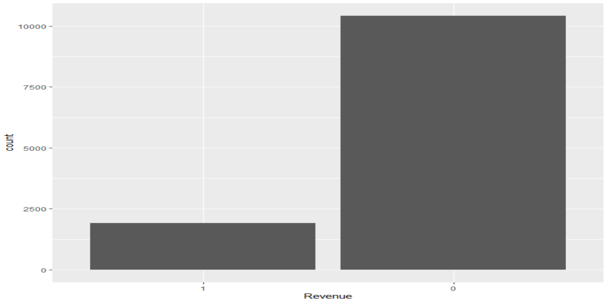

Fig. 13. Revenue class imbalance

3) Data Exploration and Transformation: The dataset was further explored to gain more insights. From Fig 14, it can be seen that the month of November had the most visits that were converted to revenue while May had the most visits that were

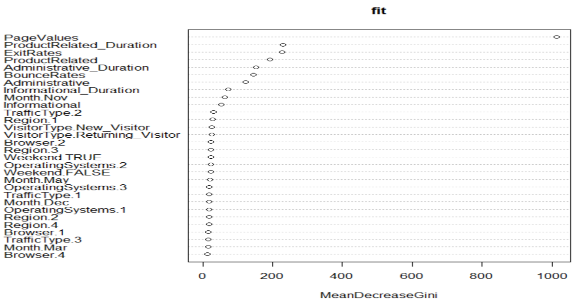

Fig. 14. Feature Selection Plot

not converted to revenue. The months May and November also had the highest amount of website visits with May being higher. Fig 15 shows a boxplot mapping the exit rates to whether the visitor visited during the weekend or not and the type of visitor. It is seen that returning visitors have lesser exit rates during the weekends while new visitors had slightly higher exit rates. It is also seen that returning visitors have higher exit rates than new visitors.

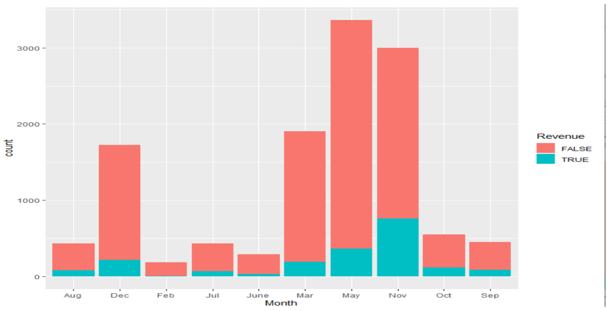

Fig. 15. Revenue by Month

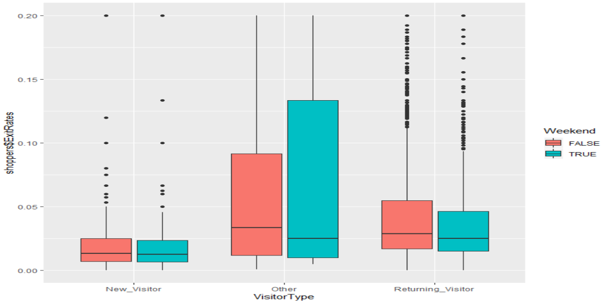

Fig. 16. Exit Rates by Weekend and Visitor type

4) Data Mining: The data mining algorithms employed for the classification are logistic regression and SVM. Due to the data imbalance, the training data set was upsampled to avoid model bias towards one class. For the SVM algorithm, two different kernels, linear and radial basis were used and the caret package was used to automate the tuning process of the cost function parameter to get the best model. The linear kernel SVM had an accuracy of 88.4 percent and a kappa value of 0.6077 showing moderate agreement between the model's predictions and the actual values. The radial basis kernel SVM had an accuracy of 87.19 percent and a kappa value of 0.5814 while, the logistic regression algorithm had an accuracy of 85.71 percent and a kappa value of 0.548.

5) Interpretation and Evaluation: The confusion matrix of the models are tabulated as follows.

TABLE VI
SVM LINEAR KERNEL CONFUSION MATRIX

	Predicted(1)	Predicted(0)
Actual(1)	462	293
Actual(0)	147	2892

TABLE VII
SVM RADIAL BASIS KERNEL CONFUSION MATRIX

	Predicted(1)	Predicted(0)
Actual(1)	468	345
Actual(0)	141	2840

TABLE VIII
LOGISTIC REGRESSION CONFUSION MATRIX

	Predicted(1)	Predicted(0)
Actual(1)	468	401
Actual(0)	141	2784

In an ecommerce scenario, identifying an opportunity for a sale from a visit is the very basis of building the models hence from the dataset, 1 is the positive class and 0 is the negative class. The models should as much as possible, be able to minimize potential sales that are mis-classified as it leads to loss in revenue. That is, the model should minimize false negatives as much as possible. This makes specificity an important evaluation metric for the models. From the results, it can be seen that the SVM linear model outperformed the other algorithms in terms of accuracy and kappa values. It also minimized the false negatives the least as is reflected in the specificity scores making it the best model amongst the three models considered. This is also reflected in the AUC as the linear kernel had the highest value.

TABLE IX
MODEL COMPARISON

Model	Accuracy	Kappa	Specificity	sensitivity	AUC
SVM linear	88.4	0.6077	0.9080	0.7586	0.7818
SVM radial	87.19	58.14	0.8917	0.7685	0.7642
Logistic	85.71	0.548	0.8741	0.7685	0.7452

V. CONCLUSION

This research applied, evaluated and compared machine learning algorithms across 3 domains, housing, ecommerce and credit default. 3 regression algorithms multiple linear regression, bagged random forest regression and decision tree regression were used to predict house rental prices and their performances evaluated based on the MSE, RMSE and MAE. The bagged random forest algorithm performed better than the other two algorithms as its predicted values were closer to the actual values hence lower MSE, RMSE and MAE values. For credit default, two algorithms, KNN and decision tree classification were used to predict the likelihood of defaulting on a credit card based on a some features. They were evaluated based on accuracy, kappa, specificity and sensitivity. The decision tree algorithm had lower accuracy and kappa scores but was able to minimize the mistakes made classifying defaulters as non defaulters as reflected in the higher sensitivity score hence making it the better model in that regard. Two algorithms SVM and Logistic regression were applied for prediction of the likelihood of a website visiting customer to make a purchase. The SVM was built using two kernels, linear and radial basis. The SVM linear model outperformed the radial basis kernel and the logistic regression algorithm as it had higher accuracy and kappa values. It was also able to make less mistakes classifying potential buyers as non buyers as reflected in the highest specificity score. This research further enhances the point that machine one machine learning algorithm's superiority over the other is highly case dependent and based on the domain of interpretation as using just evaluation metric scores without domain knowledge can be misleading. In the analysis, the models performed well but improvements can be made using advanced techniques such as ensembles and boosted algorithms.

For future research, these models can be improved and applied in other domains of research to judge their performance.

References

[1] B. Lantz, *Machine learning with R*, 3rd ed. Packtz, Birmingham UK, 2019.

[2] G. James, D. Witten, T. Hastie and R. Tibshirani,An Introduction to Statistical Learning, with applications in R, 3rd ed., USA: Springer, 2013. [Online]. Available: https://www.pdfdrive.com/an-introduction-to-statistical-learning-e29167659.html [Accessed on: October 28, 2019].

[3] Y. Sayjadah, I. A. T. Hashem, F. Atolaibi, and K. A. Kasmiran, "Credit crad default prediction using machine learning techniques",*in 2018, 4th Int. Conf. on advances in computing, communication and automation, Subang Jaya, Malaysia, Oct. 26-28, 2018*. doi:10.1109/ICACCAF.2018.8776802

[4] S. Birla, K. Kohli, and A.Dutta, "Machine learning on imbalanced data in credit risk",*in 2016, 7th Annual information technology, electronics and mobile communication, Vancouver, BC, Canada, Oct. 13-15, 2016*. doi:10.1109/IEMCON.2016.7746326

[5] R. E. Turkson, E. Y. Baagyere, and G. E. Wenya, "A machine learning approach for predicting bank credit worthiness,"in 2016, 3rd Int. Conf. on artificial intelligence and pattern recognition, Lodz, Poland, Sept. 19-21, 2016. doi:10.1109/ICAIPR.2016.7585216

[6] V. L. Kumar, S. Natarajan, S. Keerthana, K. M. Chinmayi, and N. lakshmi, "Credit risk analysis in peer-to-peer lending system,"*in 2016, IEEE Int. Conf. on knowledge engineering and applications, Singapore, Singapore, Sept. 28-30, 2016*. doi:10.1109/ICKEA.2016.7803017

[7] D. Banerjee, and S. Dutta, "Predicting the house price direction using machine learning techniques,"in 2017, IEEE Int. Conf. on Power, control, signals and instrumentation engineering, Chennai, India, Sept. 21-22, 2017. doi:10.1109/ICPCSI.2017.8392275

[8] D. Feng, Z. Liu, X. Wang, Y. Chen, J. Chang, D. Wei, anad Z. Jiang, "Machine learning based compressive strength prediction for concrete: an adaptive boosting approach", *Construction and building materials*, vol. 230, pp. 117000-117010, Jan. 2019. [Online]. Available:https://www.sciencedirect.com/science/article/pii/S095006 1819324420 [Accessed on: Decmber 1, 2019]

[9] K. Kim, J. Y. Choi, T. K. Yoo, S. K. Kim, K. Chung, and D. W. Kim, "Mortality prediction of rats in acute hermorrhagic shock using machine learning techniques", *International Federation for Medical and Biological Engineering*, vol. 51, pp. 1059-1067, June 2013.doi:10.1007/s11517-013-1091-0

[10] H. G. Santos, F. G. Zampieri, K. Normilio-Silva, G. T. da Silva, A. C. P. de Lima, A. B. Cavalcanti, and A. D. P. C. Filho, "Machine learning to predict 30-day quality-adjusted survival in critically ill patients with cancer", *Journal of critical care*, vol. 55, pp. 73-78, Feb. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0883944119 307518 [Accessed on: December 2, 2019]

[11] F. Lopes, J. Agnelo, C.A. Teixeira, N. Laranjeiro, and J. Bernadino, "Automating orthogonal defect classification using machine learning algorithms", *Future generation computer systems*, vol. 102, pp. 932-947, Jan. 2020. [Online]. Available:https://www.sciencedirect.com/science/article/pii/S01 67739X19308283 [Accessed on: November 25, 2019]

[12] C. Ma, C. Chao, and B. Cheng, "Predicting patients at risk of acute renal failure in intensive care units by using artificial intelligence tools", *International Journal of Organizational Innovation*, vol. 5, no. 2, pp. 232-247, Nov. 2012. [Online]. Available:https://ezproxy.ncirl.ie/login?url =https://search.proquest.com/docview/1095773461? accountid=103381 [Accessed on: December 13, 2019].

[13] E. Entekhabi, M. H. Nazarpak, M. Sedighi, and A. Karzemzadeh, "Predicting degradation of genipin cross-linked gelatin scaffolds with machine learning", Materials science and engineering C, vol. 107, pp. 110362-110373, Feb. 2020. [Online] Available:https://www.sciencedirect.com/science/article/abs/pii/S09 28493119322684 [Accessed on: November 25, 2019]

[14] K. M. Babu, and J. R. Beulah, "Air quality prediction based on supervised learning methods", *International Journal of Innovative technology and Exploring Engineering*, vol. 8, no. 9S4, pp.206-212, July 2019. [Online] Available:https://pdfs.semanticscholar.org/5e49/7366216d 772149f7ef6a3693afe09c811725.pdf [Accessed on: November 20, 2019]

[15] Z. Chen, C. Li, and W. Sun, "Bitcoin price prediction using machine learning: An approach to sample dimension engineering", Journal of Computational and Applied Mathematics, vol. 365, pp. 112395-112408, Feb. 2020. [Online] Available:https://www.sciencedirect.com/science/article/abs/pii/S03 7704271930398X [Accessed on: November 20, 2019]

[16] M. G. Signorini, N. Pini, A. Malovini, R. Bellazzi, and G. Magenes, "Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring", Computer Methods and Programs in Biomedicine, vol. 185, pp. 105015-105025, Mar. 2020, [online] Available:https://www.sciencedirect.com/science/article/abs/pii/S0169 260719308107 [Accessed on: November 20, 2019]

[17] S. H. Pour, A. K. A. Wahab, and S. Shahid,"Physical empirical models for prediction of seasonal rainfall extremes of peninsular Malaysia",*Atmospheric Research*, vol. 233, pp. 104720-104735, Mar. 2020, [Online] Available:https://www.sciencedirect.com/science/article/pii/S01698 09519311469 [Accessed on: November 20, 2019]

[18] Datafiniti, "Vacation rental properties in palm springs,"2016. [Online]. Available: https://www.kaggle.com/datafiniti/palm-springs-vacation-rentals [Accessed on: Nov. 20, 2019]

[19] Yeh, I. C., Lien, C. H. "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients", *Expert Systems with Applications*, vol.36, no. 2, pp. 2473-2480, 2009. [Online] Available: https://bradzzz.gitbooks.io/ga-seattle-dsi/content/dsi/dsi05classificationdatabases/2.1-lesson/assets/da tasets/ DefaultCreditCardClientsyeh_2009.pdf [Accessed on: Dec. 1, 2019]

[20] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks", *Neural computing and applications*, vol. 31, pp. 6893-6908, 2019. [Online] Available:https://link.springer.com/article/10.1007%2Fs00521-018-3523-0 [Accessed on: Dec. 1, 2019]