EVALUATION OF MACHINE LEARNING ALGORITHMS FOR REGRESSION AND CLASSIFICATION PROBLEMS

OLUWATOBI EKUNDAYO
19173105
MSc. Data Analytics, National College of Ireland

Abstract — In carrying out prediction of a target variable, the data type and research question determines the range of machine learning algorithms that can be applicable. The accuracy of a model is one significant factor that is examined in prediction. This paper compares a variety of machine learning models such as Random Forest, Logistic Regression, Support Vector Machine, Decision Trees, Multiple Linear Regression and Naïve Bayes Classifier to evaluate the performance of each model in the prediction of housing price, bank loan status and crime type. The models were compared using specific evaluation metrics and the best model was recommended as first choice of model for predictive analysis. With exploration data analysis, the main characteristics of the datasets were identified. This initial process of investigation helped discover patterns and anomalies in the data. The compared machine learning models in this paper, showed that the Random Forest model performed better in prediction of the target variables for both regression problem (housing price) and classification problems (bank loan status and crime type).

Keywords: Random Forest, Regression, Classification, Performance, Machine Learning.

I. INTRODUCTION

The increase in data over the past decade as led to development of several machine learning algorithms to discover significant patterns, provide better analysis and predictions. The decision to select the optimal algorithm to solve a task from the variety of machine learning algorithms can become very cumbersome. The aim of this project is to compare a variety of machine learning techniques in their performance on both regression and classification problems. The result of this evaluation will suggest the model to be considered as first choice of model for either of the named problems. To achieve this, three datasets (Housing Price data, Bank Loan data, and Crime data) were selected, cleaned, analysed and results from selected machine learning algorithms compared for each dataset.

Housing Price data

Housing price is one of the key elements used in identifying a city's economy. Houses are an asset and people are always looking to purchase such property at the most affordable price possible. Housing price is significant to both realtors and buyers. The dataset (California housing prices) was used for regression problem. The research questions answered using this dataset are: Which model best predicts the price of houses in California? Which variable(s) is more significant in the prediction of housing prices in California?

Bank Loan data

Loan lending is quite important to individuals and organizations to overcome financial limitations. It benefits the borrowers and it is profitable to the lender. Financial lending institutions assign credit scores (a measure of a borrower's risk and credit worthiness based on historical data) to evaluate credit risk. The dataset (Bank loan status) was used for classification problem. The research questions answered using this dataset are: Which model best predicts the loan status of a customer to be fully paid or charged off? What factors predicts the likelihood of a customers' loan status to be fully paid or charged-off?

Crime data

A crime is an activity or action that violates laws and usually with a negative impact on individuals, organizations, and communities. One of the major aims of governments is to tackle, control and possibly reduce crime rate. Several studies and research work have been carried out on analysis of criminal activities, crime predictions and pattern recognition. The dataset (Crime in Los Angeles) was used for classification problem. The research questions answered using this dataset are: Which model best predicts the two major categories of crime (crime type) in Los Angeles? Which variables best predicts the crime type?

II. RELATED WORK

In the study by Byeonghwa et al [1] to develop a prediction model for housing prices, the performance of four machine learning techniques was tested by measuring how accurately a technique could predict if the closing price was less than or greater than the price listing. For all the tests carried out, the accuracy of the Ripper model out-performed other prediction models for housing price. The paper by Pan & Zhong, compared the performance of several machine learning techniques: lasso regression, random forest, logistic regression, bridge regression and neural net. Their result showed that random forest model provided a better result [2]. Artificial neural network (ANN), support vector machine (SVM) and ensemble techniques were applied by Quang-Thanh and Nhu-Hiep,

to predict sales price for residential properties. Their results showed that the ANN and SVM were a more feasible forecasting model for effective prediction [3]. G. Naga Satish et al suggested the use of lasso regression model for predicting the price of house [4]. Their suggestion was based on the higher accuracy seen with lasso regression when compared with XG Boost model. The study on housing price prediction by Nguyen An. [5] explored important attributes using linear regression, support vector regression and random forest. They also recommended three (3) websites (Zillow, Trulia and Redfine) that provide good estimates on housing valuation (price) based on characteristics of the house. For these reasons random forest, support vector machines, multiple linear regression, were selected to truly validate which model provides better result for housing prices.

In the study by Aslam et al [6], they applied logistic regression, decision trees, support vector machine and neutral networks to predicting credit risk though credit scores. Their paper advised for more research on false negatives in credit risk assessment domain. Lin Zhu et al [7] carried out "a study on predicting loan default based on random forest algorithm". Linear regression, decision tree and other machine learning models were tested with, but Random Forest algorithm outperformed other models in the prediction of default samples. Tian-Shyug et al [8] highlighted that the objective of credit scoring models was for assigning credit applicants to either a 'good credit' group or 'bad credit' group, for the purpose of identifying those who are likely to repay or default their financial obligation. Likewise, a prediction of credit score will be carried out in this paper. Malekipirbazari and Aksakalli [9] identified high-quality peer-2-peer (P2P) borrowing customers by comparing random forest model with different machine learning methods. Their results showed that the random forest model had more significant preference than the LC Grades and FICO credit scores in identifying the good borrowers. Anchal and Ranpreet [10] carried out a research to evaluate and establish the finest model to forecast the finance status for an organization. They experimented five times on the same dataset to find the accuracy for several models used. Their results showed that the Tree Model for Genetic Algorithm was the best model for forecasting the finance for costumers.

In the study carried out by Prajakta and Vaishnavi on "predictive modelling on crime dataset using data mining", [11] they aimed at predicting the highest influencers affecting the high crime rate by using accurate predictive models. The concluded prediction was to provide the Chicago police more insights to take necessary actions. Three predictive models were investigated. A general framework was proposed by Chen Hsinchun et al [12] for crime data showing relationships between the crime types at local, national, and international levels and data mining methods applied in criminal and intelligence analysis. Ozgul Fatih et al [13] carried out studies that showed the sources of crime and with data mining pointed out which forms of knowledge discovery was appropriate for the chosen methodology. Al Boni et al carried out works on Area-Specific Crime Prediction Models [14], for which crimes were predicted by using zip code located in specific areas. In a paper by Ratul, Md, and Aminur Rab., various classification algorithms such as random forest, decision tree, adaboost classifier, extra tree classifier, linear discriminant analysis, k-neighbors classifiers, and 4 ensemble models were implemented to classify fifteen different classes of crimes [15]. According to their study, all models tested with, exhibited satisfactory accuracy except for adaboost Classifier.

III. METHODOLOGY

The methodology used for this research paper is the Knowledge Discovery in Database (KDD). The process to finding and evaluating patterns, involved the application of the following process: data selection, data pre-processing, data exploration and transformation, data mining, and data interpretation or evaluation, until knowledge was obtained from the evaluated models. The models selected for analysis in this paper are solely for understanding the relation between a given set of features and a target value (supervised learning) and are listed below:

a. **Random Forest:** This model is a powerful method for constructing a forest of random decision trees as it corrects the decision trees' habit of overfitting the training dataset [16]. It is used for classification and regression problems.
b. **Support Vector Machines:** This model uses a technique called the kernel trick [20] to transform data, upon which an optimal boundary (hyperplane) between the possible outputs that is found. It can be used for classification, regression, density estimation etc [18].
c. **Decision Trees:** This model follows a tree like architecture that simulates the decision process given a previous decision [16]. It simply starts with a root node and gradually builds "sub-trees" with internal nodes that are connected by emanating branches and ends with terminal nodes.
d. **Multiple Linear Regression:** This model assumes an approximately linear relationship among the independent variables (two or more) and the dependent variables [19]. These models are used for data with continuous quantities
e. **Naïve Bayes Classifier:** This model relies on a group of probabilistic equations based on Bayes theorem ($P(A|B)=P(B|A)*P(A)/P(B)$), which assumes that there is independence among features with the ability to consider several attributes. Hence the reason it is called naïve [23].
f. **Logistic Regression:** This model is most applicable for a dataset who's dependent or target variable is dichotomous, and its independent variables are either categorical or continuous [17].

R Programming language was used to carry out the statistical computing and data modelling for this research.

Outlined below are the steps taken to successfully carryout the analysis on each dataset.

1. HOUSING PRICE DATA

a) Data Selection

This dataset has 20640 observations (rows) and 10 attributes (columns). It has a numeric response variable and 9 predictors. It captures information about the median house prices for California districts derived from the 1990 census.

Source:https://www.kaggle.com/camnugent/california-housing-prices

b) Data Pre-Processing

Using feature extraction, the 9 variables for prediction in the dataset was converted into 12 variables. Checks carried out to confirm missing values showed that there were 207 missing values under the variable, *total bedrooms* (a predictor variable). However, the missing values were imputed using the median of the total bedrooms. Since the target variable for this dataset is *median house value*, the attribute *total bedrooms* were dropped and replaced with calculated *mean of total bedrooms* (total bedrooms / households). Likewise, the attribute *total rooms* were dropped and replaced with calculated *mean of total rooms* (total rooms / households)

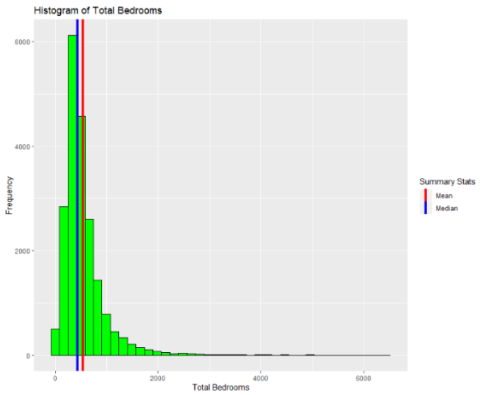

Fig. 1. Histogram plot of total bedrooms showing its mean (red line) and median (blue line)

c) Data Exploration and Transformation

Some variables required modification of its data type to the appropriate type for use. A summary of the dataset was taken, and it was observed that the *Island* category under *ocean proximity* variable had only 5, which is very low. The cleaned data finally had 20635 observations and 13 variables.

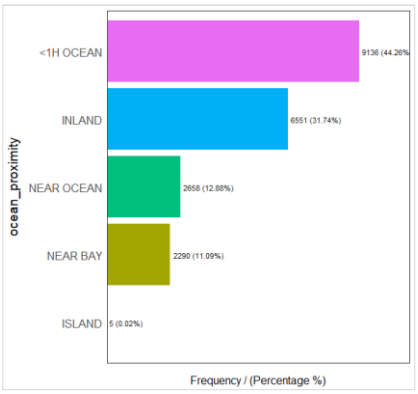

Fig. 2. Frequency Distribution of ocean proximity variable

Fig. 2. Shows the frequency of the categorical variable (*ocean proximity*). However, this variable was transformed into boolean. This is termed *one-hot encoding* in R. Also, the numerical variables were scaled so that the coefficients for models such as support vector machines are given equal weight.

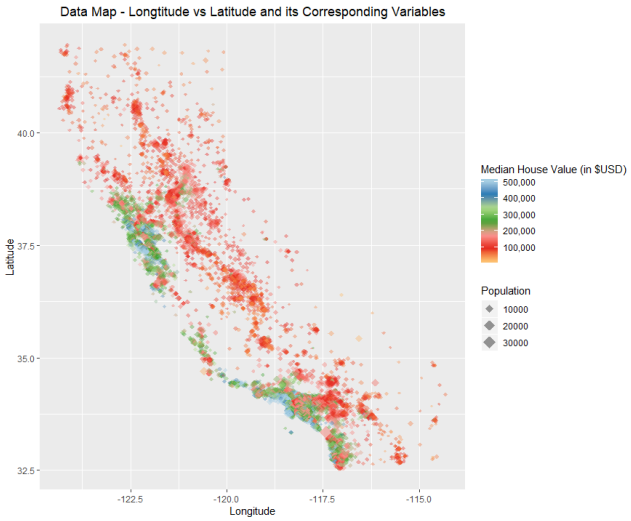

Fig 3: Map Plot

Fig. 3. shows the map plot using the *latitude* and *longitude* variable to view the frequency distribution of the median house value and population. On an average, houses nearest to the ocean (shown with the light blue colour) tends to have higher median house values, while houses with lower median values are those on the inland.

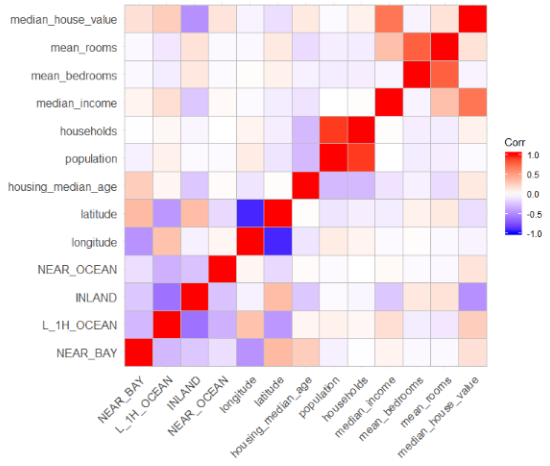

Fig 4: Correlation plot

Fig. 4. shows square colour intensity which represents the correlation. This ranges from 1 to -1. The correlation plot showed that *median income* of a buyer has the highest correlation with *median house value* (house price).

d) Data Mining

Using the train-test split method, the cleaned data was randomly split into two (ratio 70:30) with a set seed at 100. This was done so that the model could be trained and tested to avoid overfitting of the model. Four (4) models were applied to the prediction of this regression problem. The

models used are random forest, support vector machines, multiple linear regression and decision trees.

Fig. 5a, 5b, 5c and 5d below shows the summary of the four models applied for this regression problem.

```
> rfmodel_house

Call:
 randomForest(formula = median_house_value ~ ., data = train_house,      importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 2488403762
                    % Var explained: 81.33
```

Fig. 5a: Summary of Random Forest Model

```
> summary(svmmodel_house)

Call:
svm(formula = median_house_value ~ ., data = train_house)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.08333333
    epsilon:  0.1


Number of Support Vectors:  10650
```

Fig. 5b: Summary of Support Vector Machine Model

```
> summary(lm_model_house)

Call:
lm(formula = median_house_value ~ . - NEAR_OCEAN - INLAND - latitude -
    longitude - population - mean_rooms - mean_bedrooms, data = train_house)

Residuals:
     Min       1Q   Median       3Q      Max
-568380  -51303  -17164   33267  468626

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       184101.2     1006.6  182.90   <2e-16 ***
NEAR_BAY           48508.6     2297.1   21.12   <2e-16 ***
L_1H_OCEAN         39967.1     1430.3   27.94   <2e-16 ***
housing_median_age 20531.5      725.2   28.31   <2e-16 ***
households         11976.7      677.6   17.67   <2e-16 ***
median_income      77220.8      674.6  114.48   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 77940 on 14438 degrees of freedom
Multiple R-squared:  0.5445,    Adjusted R-squared:  0.5443
F-statistic:  3451 on 5 and 14438 DF,  p-value: < 2.2e-16
```

Fig. 5c: Summary of Multiple Linear Regression Model

```
> dtmodel_house
CART

14444 samples
   12 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 14444, 14444, 14444, 14444, 14444, 14444,
 ...
Resampling results across tuning parameters:

  cp          RMSE       Rsquared   MAE
  0.05275571   84162.91  0.4674244  63385.95
  0.13003557   91364.27  0.3706000  70154.71
  0.30272648  100502.16  0.2956167  78502.59

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.05275571.
```

Fig. 5d. Summary of Decision Trees Model

e) Data Interpretation

After developing the models using the training set (14444 observations), a prediction was carried out using the test set (6191 observations). To evaluate the performance of the developed models, the MAE and RMSE values were calculated for each model.

Fig. 6a, 6b, 6c and 6d below shows the calculated MAE and RMSE value of the prediction on the test set for each model. From the values shown below, we can conclude that random forest could be the best model because of low root mean squared error.

```
> test_MAE_rf <- mean(abs( predHouseTest_rf - test_house$median_house_value))
> test_MAE_rf
[1] 31784.72
> test_RMSE_rf <- sqrt(test_MSE_rf)
> test_RMSE_rf
[1] 47065.85
```

Fig. 6a: MAE and RMSE (Random Forest Model)

```
> test_MAE_svm <- mean(abs( predHouseTest_svm - test_house$median_house_value))
> test_MAE_svm
[1] 36195.28
> test_RMSE_svm <- sqrt(test_MSE_svm)
> test_RMSE_svm
[1] 54529.85
```

Fig. 6b: MAE and RMSE (Support Vector Machine Model)

```
> test_MAE_lm <- mean(abs(predHouseTest_lm - test_house$median_house_value))
> test_MAE_lm
[1] 55888.28
> test_RMSE_lm <- sqrt(test_MSE_lm)
> test_RMSE_lm
[1] 75107.09
```

Fig. 6c: MAE and RMSE (Multiple Linear Regression Model)

```
> test_MAE_dt <- mean(abs( predHouseTest_dt - test_house$median_house_value))
> test_MAE_dt
[1] 64235.99
> test_RMSE_dt <- sqrt(test_MSE_dt)
> test_RMSE_dt
[1] 84848.53
```

Fig. 6d: MAE and RMSE (Decision Tree Model)

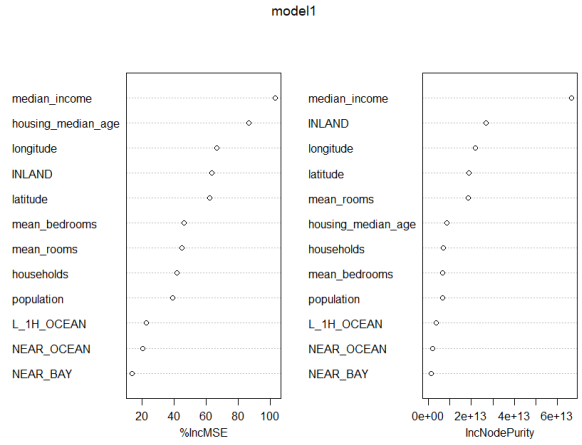

Fig. 7. Plot of Variable Importance

Percentage included mean squared error (%IncMSE) is the measure of the increase in mean squared error of predictions when the given variable is shuffled. It acts as a metric of that given variable's importance in the performance of the model.

Fig. 7. shows the plot of the important variable in the prediction. The increase in MSE of prediction (estimated with out-of-bag-CV) for random forest model shows that *median income* is an important factor for predicting housing price. This therefore answers the subset research question for this dataset.

2. BANK LOAN DATA

a) Data Selection

This dataset contains a training dataset of 100514 observations (rows) and 19 attributes (columns). It has a categorical response variable as well as several numeric predictors.

Source: https://www.kaggle.com/zaurbegiev/my-dataset

b) Data Pre-Processing

Using feature selection some attributes selected, while other attributes such as *customers-Id*, *Loan-Id*, and *month since last delinquent* were removed from the dataset. The observations with null and missing values variables such as (*bankrupt*, *years and maximum*) were removed. The attribute (*years in current job*), a 10-level categorical variable was grouped into three categories. This helped to fit the dataset better for predictive analysis.

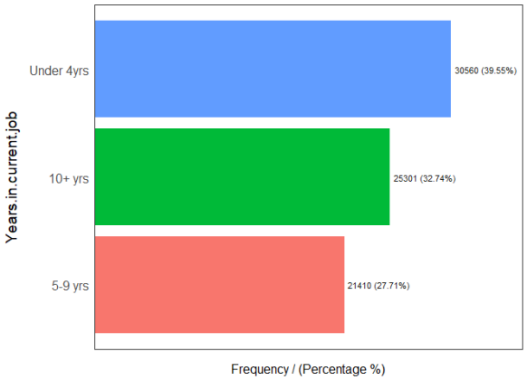

Fig. 8. Plot of customers years in current job

Fig. 8. Shows the frequency distribution of the years spent on current job for the sampled customers. Customers with a maximum of four years spent on their current job formed the largest part of the sample data (39.55%) for bank loan. 32.74% of the customers had at least 10 years in their current job, while 27.71% of the customers for bank loan had between 5 to 9 years years in the current job. Although, the variable class is unevenly distributed, the F1 Score will serve as the major metric for evaluating the performance of the applied models.

c) Data Exploration and Transformation

The summary and structure of the dataset was examined, and the data type of some variables were modified to its proper type. Categorical variables such as *home ownership* and *years in current job* were transformed using one-hot encoding in R and saved as factor data type.

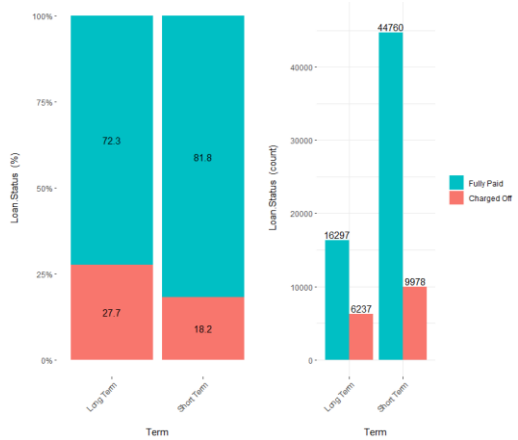

Fig. 9. Plot of Loan Term over Loan Status

Fig. 9. Shows that the likelihood of loan status being fully paid for customers who loan for short term is 81.8%, while for a long term it is 72.3%.

d) Data Mining

Using the train-test split method, the cleaned data was randomly split into two (ratio 70:30) with a set seed at 120. This was done so that the model could be trained and tested to avoid overfitting of the model. Five (5) models were applied to the prediction of this classification problem. The models used are random forest, support vector machines, decision trees, naïve bayes classifier and logistic regression.

Summary of the top three (3) models in performance are shown below.

```
> rfmodel_loan

Call:
 randomForest(formula = Loan.Status ~ ., data = train_loan, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 15.32%
Confusion matrix:
            Charged Off Fully Paid class.error
Charged Off        3071        8234 0.728350287
Fully Paid           52       42732 0.001215408
```

Fig. 10a: Summary of Random Forest Model

```
> dtmodel_loan
CART

54089 samples
   14 predictor
    2 classes: 'Charged Off', 'Fully Paid'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 54089, 54089, 54089, 54089, 54089, 54089, ...
Resampling results across tuning parameters:

  cp           Accuracy   Kappa
  0.0002192213  0.8173007  0.3515079
  0.0002380117  0.8217420  0.3563495
  0.2656085584  0.8161616  0.1719497

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.0002380117.
```

Fig. 10b. Summary of Decision Tree Model

```
> summary(svmmodel_loan)

Call:
svm(formula = Loan.Status ~ ., data = train_loan)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  20331

 ( 11903 8428 )


Number of Classes:  2

Levels:
 Charged Off Fully Paid
```

Fig. 10c. Summary of Support Vector Machine Model

e) Data Interpretation

After developing the models using the training set (54089 observations), a prediction was carried out using the test set (23182 observations). To evaluate the performance of the developed models; the accuracy, misclassification error, F1 score, precision and recall/sensitivity are measured for each model.

The out-of-bag error (OOB) is one where each tree sample not used in the construction of the tree becomes a test set. The OOB estimate of the random forest model is 15.32%.

```
> contrasts(cleaned_loan_data$Loan.Status)
            Fully Paid
Charged Off          0
Fully Paid           1
```

Fig. 11. Target Variable Index

Fig. 11. above shows how R computed the index for the target variable. This helped in identifying exact location for the false positive and false negative values.

Confusion Matrix is very useful in capturing what has happened in the evaluation test to provide extra details [23]. A whole range of performance measures can be obtained from this matrix, as it highlights different aspects of performance of the model. Confusion matrix for the five models used are shown below.

```
> # Check confusion Matrix
> tabloan_rf <- table(Actual = test_loan$Loan.Status,
 Predicted = predLoanTest_rf)
> tabloan_rf
            Predicted
Actual       Charged Off Fully Paid
  Charged Off        1287       3523
  Fully Paid           15      18357
```

Fig. 12a Random Forest Model

Fig. 12a. showed that random forest model correctly predicted 1287 bank loan status to be charged off and 18357 bank loan status to be fully paid. However, the model wrongly predicted 15 bank loan status to be charged off and 3523 bank loan status to be fully paid.

```
> # Check Confusion Matrix
> tabloan_svm <- table(Actual = test_loan$Loan.Status,
 Predicted = predLoanTest_svm)
> tabloan_svm
            Predicted
Actual       Charged Off Fully Paid
  Charged Off        1266       3544
  Fully Paid            0      18372
```

Fig. 12b. Support Vector Machine Model

Fig. 12b. showed that support vector machine model correctly predicted 1266 bank loan status to be charged off and 18372 bank loan status to be fully paid. However, the model wrongly predicted 3544 bank loan status to be fully paid.

```
> # Check confusion Matrix
> tabloan_dt <- table(Actual = test_loan$Loan.Status,
 Predicted = predLoanTest_dt)
> tabloan_dt
            Predicted
Actual       Charged Off Fully Paid
  Charged Off        1265       3545
  Fully Paid            0      18372
```

Fig. 12c. Decision Tree Model

Fig. 12c. showed that decision tree model correctly predicted 1265 bank loan status to be charged off and zero 18372. However, the model wrongly predicted 3545 bank loan status to be fully paid.

```
> # Check Confusion Matrix
> tabloan_nb <- table(Actual = test_loan$Loan.Status,
 Predicted = predLoanTest_nb)
> tabloan_nb
            Predicted
Actual       Charged Off Fully Paid
  Charged Off        3518       1292
  Fully Paid        12174       6198
```

Fig. 12d. Naïve Bayes Classifier Model

Fig. 12d. showed that naïve bayes classifier model correctly predicted 3518 bank loan status to be charged off and 6198 bank loan status to be fully paid. However, the model wrongly predicted 12174 bank loan status to be charged off and 1292 bank loan status to be fully paid. This model performed poorly for this dataset.

```
> # Check Confusion matrix
> tabloan_lr <- table(Actual = test_loan$Loan.Status,
 Predicted = predLoanTest_lr)
> tabloan_lr
            Predicted
Actual           0     1
  Charged Off 1264  3546
  Fully Paid     0 18372
```

Fig. 12e. Logistic Regression Model

Fig. 12e. showed that logistic regression model correctly predicted 1264 bank loan status to be charged off and 18372 bank loan status to be fully paid. However, the model wrongly predicted 3546 bank loan status to be fully paid.

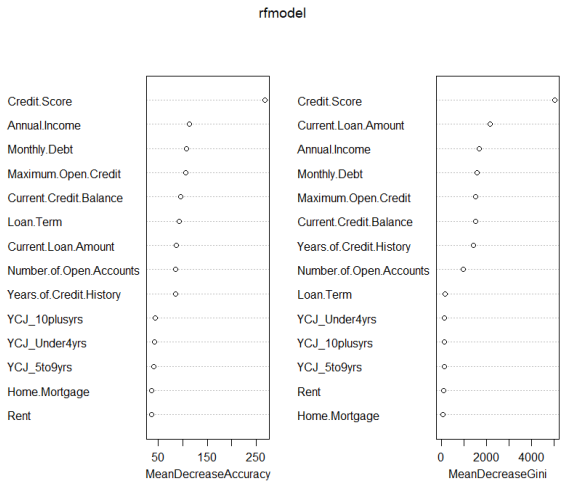

Fig. 13. Plot of Variable Importance

Mean Decrease in Gini is an effective measure of how important a variable is for estimating the value of the target variable across all the decision trees in random forest model. It is the average of a variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each individual decision tree in the random forest.

Fig. 13. shows the plot of the important variable(s) in the prediction. From the Mean Decrease in Gini plot, we can see that the customer's credit score has the highest impact in explaining the *Loan Status* from the random forest model. Customer's *credit score* is therefore the most important factor for predicting customers loan status. This therefore answers the subset research question for this dataset.

3. CRIME DATA

a) Data Selection

This dataset has 1048575 observations (rows) and 28 attributes (columns). This dataset captures incidents of crime in the City of Los Angeles from 2010 - 2019.
The target variable (*Crime Type*) is categorised into two: Violent Crime (represented as 1) and Property Crime (represented as 2).
Source: https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-2019/63jg-8b9z

b) Data Pre-Processing

Due to the large size of the data, by using a cluster random sampling method, specific groups were selected. This method involves taking every member from some of the groups. Three areas (Central, Pacific and West Valley) and three victim descents (White, Hispanics and Black) were selected. Also, using feature selection, some attributes were retained while others (without potential use or impact) were exempted from the cleaned data.

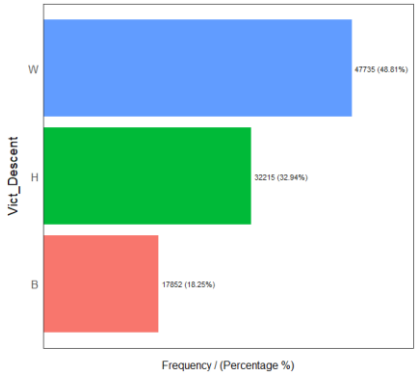

Fig. 14. Distribution of the victims descent

Fig. 14. depicts the distribution of victim's descent of the sampled data, which is composed of Whites (48.81%), Hispanics (32.94%) and Blacks (18.25%)

c) Data Exploration and Transformation

To make the data fit for the models to be used, the categorical variables with three levels were transformed using on-hot encoding. The cleaned and transformed data finally had 97802 observations and 12 variables.

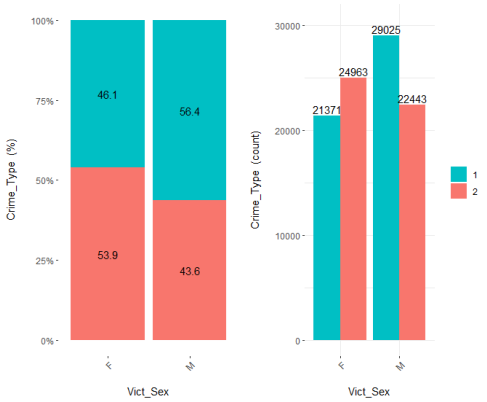

Fig. 15. Distribution of victims sex

Fig. 15. Shows that the likelihood of having violent crime on a male victim is loan status 56.4% and 46.1% for a female victim. The likelihood of having property crime on male victim is 43.6% and 53.9% for a female victim.

d) Data Mining

Using the train-test split method, the cleaned data was randomly split into two (ratio 70:30) with a set seed at 150. This was done so that the model could be trained and tested to avoid overfitting of the model. Five (5) models were applied to the prediction of this classification problem. The models used are random forest, support vector machines, naïve bayes, decision trees, and logistic regression.

Summary of the top three models in performance are shown below.

```
> rfmodel_crime

Call:
 randomForest(formula = Crime_Type ~ ., data = train_crime, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 39.72%
Confusion matrix:
      1     2 class.error
1 25334 10001   0.2830338
2 17194 15932   0.5190485
```

Fig. 16a. Summary of Random Forest Model

```
> summary(svmmodel_crime)

Call:
svm(formula = Crime_Type ~ ., data = train_crime)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  59008

 ( 29513 29495 )


Number of Classes:  2

Levels:
 1 2
```

Fig. 16b. Summary of Support Vector Machine Model

```
> summary(nbmodel_crime)
          Length Class  Mode
apriori    2     table  numeric
tables    11     -none- list
levels     2     -none- character
isnumeric 11     -none- logical
call       4     -none- call
```

Fig. 16c. Summary of Naïve Bayes Model

e) Data Interpretation

The models were developed using the training set (68461 observations), and a prediction was carried out using the test set (29341 observations). To evaluate the performance of the developed models; the accuracy, misclassification error, F1-score, precision and recall/sensitivity are measured for each model. The OOB estimate of the random forest model is 39.72%.

```
> contrasts(cleaned_crime_data$Crime_Type)
  2
1 0
2 1
```

Fig. 17: Target Variable Index

Fig. 17. above shows how R computed the index for the target variable. This helped in identifying exact location for the false positive and false negative values.

Confusion matrix for the five models used for this classification problem are shown below.

```
> # Check confusion Matrix
> tabcrime_rf <- table(Actual = test_crime$Crime_Type,
 Predicted = predCrimeTest_rf)
> tabcrime_rf
      Predicted
Actual     1     2
     1 10725  4336
     2  7342  6938
```

Fig. 18a. Random Forest Model

Fig. 18a. showed that random forest model correctly predicted 10725 actual violent crime and 6938 property crime cases. However, the model wrongly predicted 7342 violent crime caases and 4336 property crime cases.

```
> # Check Confusion matrix
> tabcrime_svm <- table(Actual = test_crime$Crime_Type,
 Predicted = predcrimeTest_svm)
> tabcrime_svm
      Predicted
Actual     1     2
     1 10578  4483
     2  8069  6211
```

Fig.18b. Support Vector Machine Model

Fig. 18a. showed that support vector machine model correctly predicted 10578 actual violent crime cases and 6211 property crime cases. However, the model wrongly predicted 8069 violent crime cases and 4483 property crime cases

```
> # Check Confusion Matrix
> tabcrime_nb <- table(Actual = test_crime$Crime_Type,
 Predicted = predcrimeTest_nb)
> tabcrime_nb
      Predicted
Actual     1     2
     1 11111  3950
     2  8877  5403
```

Fig. 18c. Naïve Bayes Classifier Model

Fig. 18a. showed that naïve bayes classifier model correctly predicted 11111 actual violent crime cases and 5403 actual property crime cases. However, the model wrongly predicted 8877 violent crime cases and 3950 property crime cases

```
> # Check Confusion Matrix
> tabcrime_dt <- table(Actual = test_crime$Crime_Type,
 Predicted = predCrimeTest_dt)
> tabcrime_dt
      Predicted
Actual     1     2
     1 12118  2943
     2  9955  4325
```

Fig. 18d. Decision Tree Model

Fig. 18d. showed that decision tree model correctly predicted 12118 actual violent crime cases and 4325 property crime cases. However, the model wrongly predicted 9955 violent crime cases and 2943 property crime cases

```
> # Check Confusion matrix
> tabcrime_lr <- table(Actual = test_crime$Crime_Type,
 Predicted = predcrimeTest_lr)
> tabcrime_lr
      Predicted
Actual    0    1
     1 9618 5443
     2 7129 7151
```

Fig. 18e. Logistic Regression Model

Fig. 18e. showed that logistic regression model correctly predicted 9618 actual violent crime cases and 7151 property crime cases. However, the model wrongly predicted 7129 violent crime cases and 5443 property crime cases

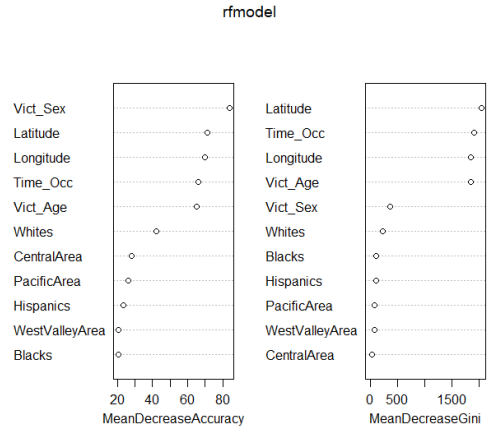

Fig. 19. Plot of Variable Importance

Fig. 19. shows the plot of the important variables in the prediction. From the Mean Decrease in Gini plot, we can see that the four variables are important in predicting *Crime Type (Violent Crime or Property Crime)* from the random forest model. The important factors for predicting crime type are *Latitude, Time of Occurrence, Longitude and Victims Age*. This therefore answers the subset research question for this dataset.

IV. EVALUATION METHODS

Evaluation metrics are excellent way to carry out performance measurement of a model. Highlighted below are comparisons of the evaluation metrics for the models used in this research on each dataset.

1. HOUSING PRICE DATA

TABLE I
EVALUATION RESULTS - MODEL COMPARISON

Machine Learning Model	MAE	RMSE
Random Forest	31784.72	47065.85
Support Vector Machine	36195.28	54529.85
Multiple Linear Regression	55888.28	75107.09
Decision Tree	64235.99	84848.53

Table I and Fig. 20. compares the evaluation metrics used for the four models applied for the regression problem.

Both MAE and RMSE values were lower for random forest model, indicating that random forest model performed better than the other three models used.

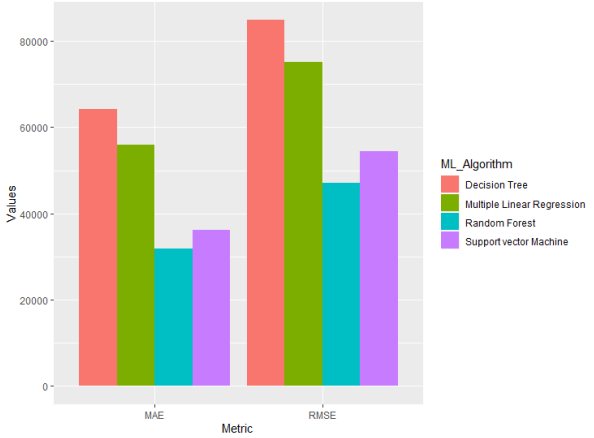

Fig. 20. Evaluation Metrics Comparison Bar Chart (Regression)

Random Forest model was able to predict the median housing price in given neighbhood to be within $47,100 of the actual median housing price in california.

To answer the research question for the classication problem, Random Forest Model best predicts Housing Price in California.

2. BANK LOAN DATA

TABLE II
EVALUATION RESULTS - MODEL COMPARISON

ML Model	F1 Score	Precision	Sensitivity
Random Forest	0.91210	0.83899	0.99918
Support Vector Machine	0.91203	0.83829	1.00000
Decision Tree	0.91201	0.83825	1.00000
Logistic Regression	0.91199	0.83822	1.00000
Naïve Bayes Classifier	0.47931	0.82750	0.33736

Table II and Fig. 21. compares the evaluation metrics used for the five models applied for the classification problem.

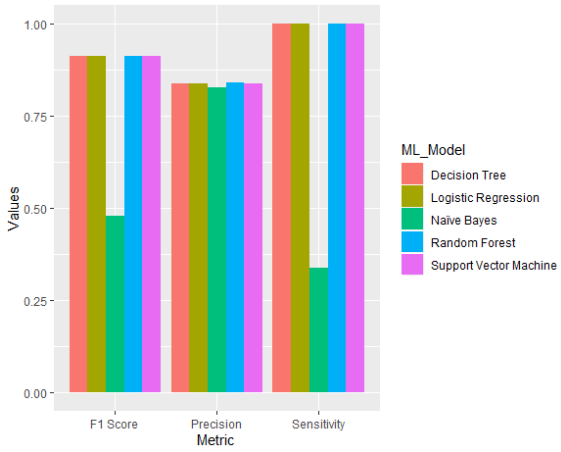

Fig. 21. Evaluation Metrics Comparison Bar Chart (Classification)

The performance of top three models (random forest, support vector machine and decision tree) are very close. However, random forest slightly had the highest F1-Score.

Table III below shows the accuracy and misclassification error comparison for the five models used.

In terms of misclassification error, random forest model had the lowest value, which also supports the performance of the model.

TABLE III
EVALUATION RESULTS - MODEL COMPARISON

Machine Learning Model	Accuracy	Misclassification Error
Random Forest	0.84738	0.15261
Support Vector Machine	0.84712	0.15287
Decision Tree	0.84707	0.15292
Naïve Bayes Classifier	0.41911	0.58088
Logistic Regression	0.00000	0.15296

For the domain of this classification problem (Bank Loan), the cost associated with false negative is very high, as the bank would not want to lose any money. Hence why the F1-Score and Sensitivity are taken into consideration.

To answer the research question for this classication problem; Random Forest model best predicts the loan status of customers of the bank.

3. CRIME DATA

TABLE IV
EVALUATION RESULTS - MODEL COMPARISON

ML Model	F1 Score	Precision	Sensitivity
Random Forest	0.54301	0.61540	0.48585
Logistic Regression	0.53219	0.56781	0.50077
Support Vector Machine	0.49740	0.58079	0.43494
Naïve Bayes Classifier	0.45724	0.57768	0.37836
Decision Tree	0.40143	0.59507	0.30287

Table IV and Fig. 22. compares the evaluation metrics used for the five models applied for the classification problem.

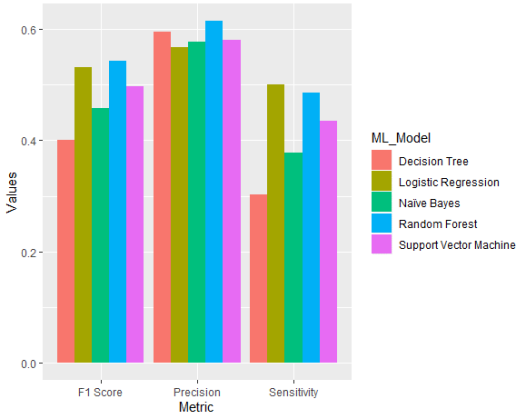

Fig. 22. Evaluation Metrics Comparison Bar Chart (Classification)

Random forest can be seen to have performed better than other models in terms of F1-Score and Precision.

Table V below shows the accuracy and misclassification error comparison for the five models used.

In terms of misclassification error, random forest model had the lowest value, which also supports the performance of the model.

TABLE V
EVALUATION RESULTS - MODEL COMPARISON

ML Model	Accuracy	Misclassification Error
Random Forest	0.6019904	0.3980096
Support Vector Machine	0.5722027	0.4277973
Naïve Bayes Classifier	0.5628302	0.4371698
Decision Tree	0.5604103	0.4395897
Logistic Regression	0.1855083	0.4284789

For the domain of this classification problem (Crime), the cost associated with false positive is high. If precision is not high, less priority and attention might be given to a specific type of crime. Hence why the F1-Score and Precision are taken into consideration.

To answer the research question for this classication problem; Random Forest model best predicts the 2 types of crime in Los Angeles.

Evaluation metrics used for the classification problems are defined below:

Root Mean Square Error (RMSE): is the square root of the average of squared differences between the predicted and actual observation.

Mean Absolute Error (MAE): is the average of absolute differences between prediction and actual observation where all individual differences have equal weight.

F1-Score or F Measure: is the weighted average of Precision and Recall/Sensitivity. This score takes both false positives and false negatives into account. In the case of an uneven class distribution, F1 Score is usually more useful than accuracy [23].

F1 Score = (2 x Precision x Recall) / (Precision + Recall)

Accuracy: is a ratio of correctly predicted observation to the total observations. Accuracy is the most intuitive performance measure.

Accuracy = (True Positive + True Negative) / Total Population

Precision: shows how often the prediction is correct when a positive value is predicted

Precision = True Positive / (True Positive + False Positive)

Recall/Sensitivity: shows how often the prediction is correct when the actual value is positive.

Recall = True Positive / (True Positive + False Negative)

V. CONCLUSION AND FUTURE WORK

It can be concluded from the predictive analysis carried out in this paper that Random Forest is a better model for both regression and classification problem. This however is subjected to further test with datasets with an even class distribution. Random forest model is a much simpler model to run due to less hyperparameter tuning. Random Forest model is therefore recommended to be considered as first choice of model for carrying out predictions on both regression and classification problems.

Given more time, hyperparameters for the models would have been tuned and results compared. Also, boosting techniques would be tested and compared for both regression and classification problems.

Further research is recommended with multicollinear datasets with various number of variables for the regression problem. I suggest other models such as extreme gradient boosting to be tested. Also, to make better predictions of housing price, other variables such as measurable housing features can be used. For the classification problems, an evenly distributed dataset should be tested and with application of other bagging and boosting algorithms.

VI. REFERENCES

[1] P. Byeonghwa and J. Kwon Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data", Expert Systems with Applications 42, 2015.

[2] P. Shuaidong and Z. Jianyuan. House Pricing Prediction Report, 2019.

[3] B. Quang-Thanh and D. Nhu-Hiep, "House Price Estimation in Hanoi using Artificial Neural Network and Support Vector Machine: in Considering Effects of Status and House Quality", 2017

[4] G. Satish, V. Raghavendran, M. Sugnana Rao and Ch. Srinivasulu, "House Price Prediction Using Machine Learning", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8, Issue-9, July 2019.

[5] An. Nguyen, "Housing Price Prediction," 2018.

[6] A. Uzair, A. Tariq, S. Asim and B. Nowshath, "An Empirical Study on Loan Default Prediction Models," Journal of Computational and Theoretical Nanoscience, Volume 16, 3483-3488, 2019.

[7] L. Zhu, D. Qiu, D. Ergu, C. Ying and K. Liu, "A study on predicting loan default based on the random forest algorithm," 7th International Conference on Information Technology and Quantitative Management, 2019.

[8] L. Tian-Shyug and I. Chen, "A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines", Expert Systems with Applications, 28, pp. 743 – 752, 2015.

[9] M. Malekipirbazari and V. Aksakalli, "Risk assessment in social lending via random forests," Expert Systems with Applications, 42, 2015, pp. 4621-4631.

[10] A. Goyal, R. Kaur, "Accuracy Prediction for Loan Risk Using Machine Learning Models," International Journal of Computer Science Trends and Technology (IJCST) – Volume 4 Issue 1, Jan - Feb 2016.

[11] P. Yerpude and V. Gudur, "Predictive Modelling of Crime Dataset Using Data Mining," International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol. 7, No. 4, July 2017.

[12] H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin and M. Chau, "Crime Data Mining: A General Framework and Some Examples," IEEE, Journals and Magazines, Vol 37, Issue 4, 2004.

[13] F. Ozgul, C. Atzenbeck, A. Celik and Z. Erdem, "Incorporating data sources and methodologies for crime data mining," IEEE International Conference on Intelligence and Security Informatics, 2011.

[14] M. Boni and M. Gerber, "Area specific crime prediction models," 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 671-676, 2016.

[15] Ratul, Md, and Aminur Rab. "A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining." Vol. XIV, 2001.02802, 2020.

[16] Lior Rokach, Oded Maimon, "Data Mining with Decision Trees: Theory and Applications," Second Edition, pp. 15, 27.

[17] Sammut C., Webb G.I. (eds), "Logistic Regression In: Encyclopedia of Machine Learning and Data Mining," Springer, Boston, MA, 2017

[18] X. Zhang, C. Sammut, and G.I. Webb (eds), "Support Vector Machines In: Encyclopedia of Machine Learning and Data Mining," Springer, Boston, 2017.

[19] N. Quadrianto and W.L. Buntine, "Multiple Linear Regression. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining," Springer, Boston, MA, 2017.

[20] Ch.Sai Sindhu, T.Hema Sai, Ch.Swathi, S.Kishore Babu, "Predictive Analytics Using Support Vector Machine," International Journal for Modern Trends in Science and Technology, Vol. 03, Special Issue 02, 2017, pp. 19-23.

[21] T. Pang-Ning, M. Steinbach and V. Kumar, "Introduction to Data Mining," Pearson Addison, Wesley, 2006, pp. 187, 227, 296, 297, 729.

[22] I.H. Witten, E. Frank, M.A. Hall, Christopher J. Pal. "Data Mining, Practical Machine Learning Tools and Techniques", 2017, pp. 186-190.

[23] J.D. Kelleher, B.M. Namee and A. D'Arcy. "Fundamentals of Machine Learning for Predictive Data Analytics, Algorithms, worked examples and case studies," 2015, pp. 267, 367, 402.