APPLICATION OF MACHINE LEARNING MODELS IN DECISION SUPPORT SYSTEMS

Kirubakaran Balaraman
School of Computing
National College of Ireland
Dublin, Ireland
x19241658@student.ncirl.ie

***Abstract*—The project focuses on implementing Machine learning models on three different datasets from domains like Banking and Finance, Telemarketing and Hospitality. This involved predicting the loan status of a customer which helps in risk analysis for banks, whether a client will subscribe to a term deposit and predicting the price of Airbnb hotels. This study followed Knowledge Discovery in Database methodology (KDD) and and followed the process for data selection to knowledge discovery throughout the study.**

The research contained both classification and regression problems in which different algorithms like Random Forest, Logistic Regression, Support Vector Machines (SVM), Multiple Linear Regression and XGBoost Regression were used. Every model was trained and evaluated with various metrics and tuned to improve their performance using cross validation techniques.

The classification models were evaluated with metrics like Receiver Operating Characteristic (ROC), classification accuracy, Specificity, Sensitivity andArea Under the Curve (AUC). The regression models were evaluated by comparing Mean Square Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE)

The trained models were evaluated and they can be used in a similar problem and will classify and predict most of the cases correctly.

Keywords—Regression, Classification, Cross validation, XGBoost algorithm, Linear Regression, Logistic Regression, KDD, Support Vector Machines, Random Forest

I. INTRODUCTION

The success of every business is a step by step process and each decision made throughout the time contributes to the growth in either positive or negative direction. Decision making plays a key role in the future growth of a business. It helps in identifying goals, reduction of the resources and helps to find alternative ways to tackle a problem. This project focuses on domains where making such decisions quickly and efficiently with the help of historical data leads to profits.

Every busines make marketing campaigns to improve their business grow and find new leads and potential customers. In order for banks to sustain and continuously serve the people by providing loans and other services, they need to maintain enough liquidity. This can be achieved by means of the term deposits and fixed deposits made by the customers in the bank. The first dataset used in this project is about the marketing campaigns made by a bank to gain new customers by using telemarketing. To generate a lead or attract customers, banks cannot call everyone in their database as it would be a time consuming process and requires lot of manpower and resources. So banks have to follow some technique to reduce the number of calls made to customers without reducing the chances of getting a new customer. Managers must use a Direct Marketing strategy to target customers who have a higher chance of subscribing for term deposits such that reducing resources used like workforce and calls made. This can be achieved by analysing the historical data and finding out various factors that make a customer to subscribe. An efficient way of doing it is to use the power of machine learning algorithms to train classification models and make them able to predict with the help of various factors that has high influence on the customer's decision. In this study we use Random forest and Support vector Machine to help banks improve their direct marketing strategies and identify customer who are more likely to subscribe.

The banking sector majorly contributes to the economy of the world. Small scale Businesses, individuals and even enterprises rely on bank loans to improve their business. For a bank ro provide loan it needs enough liquidity. This is related to our previous data set in a way such that a bank gain liquidity with the deposits done and provides that amount as loan to those who need it. Before lending money to a customer, banks have to carry out risk analysis abased on various factors to find out whether the customer will repay the loan without defaults. Implementing classification algorithms to this problem helps banks to solve these complex. For this study, we use a bank loan status dataset and use classification algorithms like random forest to predict the loan status of a customer.

The hospitality industry plays a major role in the global GDP and contributes more to the economy. There are various big business groups to small hotels involved in this industry. One such big business is Airbnb which is responsible for connecting hosts and travellers thus creating a huge impact in this industry. The hotels and hosts have to offer best price to the travellers and guests without compromising the profits. They have to change the price according to time of the day, availability, demand and also based on the season to get more profits and attract customers. This is widely known as Dynamic pricing which is determining the real-time price of the hotels based on various factors. Regression techniques can be used to predict the prices based on various features and improve the way the business works and increase the profit as well as enable guests to get best deal.

The paper starts by discussing about the previous researches carried out in the following section, with the help of the same datasets and similar works. In the third section, the introduction about the methodology used and the implementation of the models that are applied to the datasets are discussed. In the section followed by that, the machine learning models are evaluated using varous metrics and compared with one another to find the best performing models.

II. Related Works

The machine learning is a vast area of study and variuous research have been conducted on similar datasets that we have used in our project and similar goal. The researches focused on applying various models for same problems and find the optimum performance by applying various techniques and tuning methods.

The research on a similar problem is conducted by S.Moro in [1] in which they analysed the same dataset and compared performance of logistic regression, decision tress (DTs) , neural network (NN) and Naïve Bayes. They used the Area under ROC curve (AUC), which was 0.8, to evaluate the model and reached a 79 % accuracy in classifying the results. They used some methods for extracting knowledge and also performed sensitivity analysis in their research.

E. Zeinulla et. al. in [2] worked on a similar problem and applied five different models namely Random Forest, Decision Trees, Artificial Neural Networks, Support Vector Machines, Logistic Regression and k-Nearest Neighbours and found random forest to be performing well. They didn't include all the features and carried out feature columns and performed series of performance tuning and obtained an accuracy of 90.884 %. They used Receiving Operator Characteristic (ROC) curve and Cumulative Accuracy Profile (CAP) to evaluate the models.

G. Arutjothi and C. Senthamarai in [3] were successful in implementing a k-Nearest Neighbour classifier on a bank dataset to analyse the credit risk of a client by predicting whether they default on loan. The features used were normalized using Min-Max Normalization to simplify the dataset and also remove the outliers. They were able to achieve a classification accuracy of 75.08 percentage by suing KNN model in many iterations. They achieve the optimum performance when the data was equally split into training and test sets.

In [4], Z. Peng et. al. worked in predicting the second hand house price for a Chinese house dataset by selecting 10 features based on importance of the features and whether they are significantly contributing in predicting the price. They applied DTs regression, Multiple Linear Regression and Extreme Gradient Boosting (XGBoost) to the dataset and found the optimum model to be XGBoost with a learning rate of 0.1 and gamma value of 0.9, which predicted with a R-squared value of 0.9251.

A Cost-Sensitive Algorithm with Random Forest model was built by D. Devi et. al. in [5] to detect fraudulent activities with credit card. This was done based on assigning weights instead of voting strategy. The dataset was highly imbalanced and predicted the built a Cost-Sensitive Random Forest algorithm by using a weighted strategy instead of a voting scheme to detect credit fraud. Using the cost sensitive approach helped with the imbalance in the dataset and outperformed the standard one. A ten fold cross validation was carried out in the study and achieved an accuracy of 82.6% where the standard RF just predicted 68.4% of the data correctly. Confusion matrix, AUC value and F score were used to evaluate the model. with imbalanced data.

In [6], G. Anuradha et. al. applied various possible regression models to predict the house price to help the sellers decide the optimum selling price and compared their performance. They used regression algorithms like Multiple Linear, LASSO, Ridge, Elastic Net, Ada Boosting and Gradient Boosting and found the latter to be the best performer. The models were evaluated by comparing the Mean Square Error (MSE) and Root Mean Square Error (RMSE). The Gradient Boosting model showed an R square of 0.917, whereas the Elastic Net showed the lowest, which is 0.66.

Y. Li, in [7] worked on building a classification model to help banks find the repayment capability of the customers before lending a loan to them. This study concentrated on only two models namely, logistic regression and XGBoost classification. The evaluation metrics used are Area under the curve (AUC) and the Population Stability Index (PSI). When the models were trained and tested , XGBoost showed better performance with AUC of 0.9079 had a good stability value of 0.0565 in the test and verification datasets.

In [8], Krichene et. al. analysed a Tuinisian commercial bank dataset to provide a machine learning solution to assess loan risk by using Naïve Bayes classifier. In this paper, 32 features were used to predict the response variable loan default. The model correctly classified 63.85 percent of the test data and it was evaluated by plotting ROC curve with an AUC value of 69 percent. A neural network was then used to classify the test dataset and it showed an accuracy of 83 percent.

The risk of employee leaving the organization is predicted by D. S. Sisodia et. al. in [9] by using classifiers like Naïve Bayes, Random Forest, K-Nearest Neighbours, Linear Support Vector Machines (SVM) and decision tree C5.0. The models were compared based on the accuracy, Recall, Specificity, Sensitivity and Precision. The ROC curve was plotted to know about the false and true positives. Based on the above mentioned metrics the researchers found that the Random Forest model performed better with a overwhelming 0.9897 accuracy value.

A. S. Galathiya et. al. worked in improvising the decision tree algorithm in [10], by making changes to the existing algorithm. The new algorithm carries out feature selection, prunes error and does cross validation. The changes to existing algorithm was done by using RGUI from weka package. This resulted in a 1-3 % increase in accuracy for every new dataset fitted using that compared to the existing decision tree algorithms like ID3, C4.5 and C5.

The paper [11] discusses on optimum feature selection for logistic regression while trying to detect pedestrians from INRIA person dataset. The model was trained in a forward selection method which is step-wise and achieved an accuracy of 95 %. Out of the classified images 10% of those were false positive which is not too high to be problematic.

In the paper [12], M. Sivasakthi used machine learning algorithms like C4.5 decision tree, Multilayer propagation and Naïve Baye to determine student's performance on programming. The precision and recall of the classification and the accuracy were used to evaluate and compare the models to find the optimum one. All models showed an accuracy over 80 percentage. But Decision tree performed well with an accuracy of 92.03 % where NB's accuracy was 84.46 %. Multilayer perception also performed well with accuracy of 90% and it is a a neural network incorporated with back propagation.

In [13], Kavitha S et. al. applied regression models like Linear Regression and Support Vector Machines (SVMs) on students assessment data from UCI repository and assessed the performance of the models with help of metrics like Mean Absolute Error (MAE) and RMSE. Linear regression was the best performer among them with lesser MAE of 9.6689 and 12.54 RMSE.

T. Lu et. al built an intrusion detection system with the power of random forest and fitted the model to the KDDCup99 dataset. A oversampling technique called Synthetic Minority Oversampling (SMOTE) used to tackle the high imbalance in the data. F1- score , precision and recall were used as the metrics of evaluation. The main focus was to improve the F1 score as there was huge imbalance and they achieved it with a score of 0.99.

III. Methodology

The project followed the Knowledge Discovery in Databases (KDD) method to meet the objectives and gain insights from the three different datasets chosen for analysis. The steps of KDD are followed from selecting data to discovering knowledge.

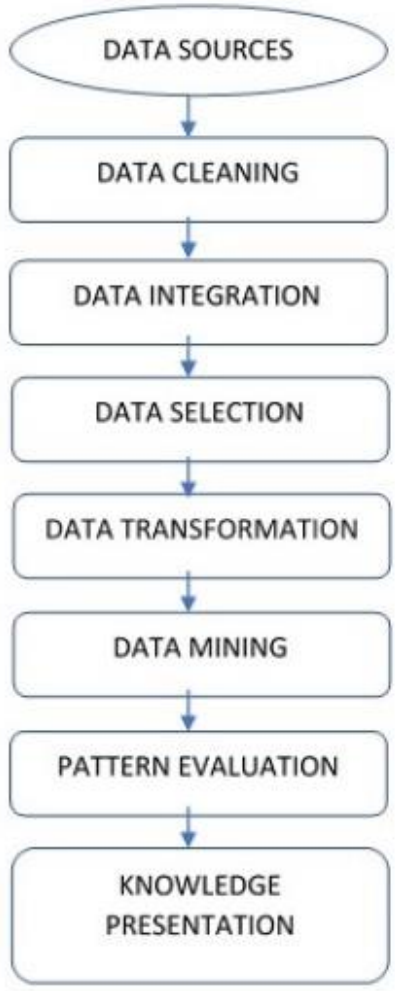

Figure 1: KDD process flow

A. Telemarketing

The classification models, Naïve Bayes and Support Vector Machines (SVM), were used to predict whether a customer will subscribe to the term deposit or not by following the below process.

i. Data Selection: The initial dataset used in this problem is sourced from UCI repository [19] and has 21 columns and 41188 rows. I selected all the columns from the raw dataset as they were relevant and formed the target data and carried out to the next stage of the process.

ii. Pre-processing and Transformation: The data was checked for null values and found to be free of it.

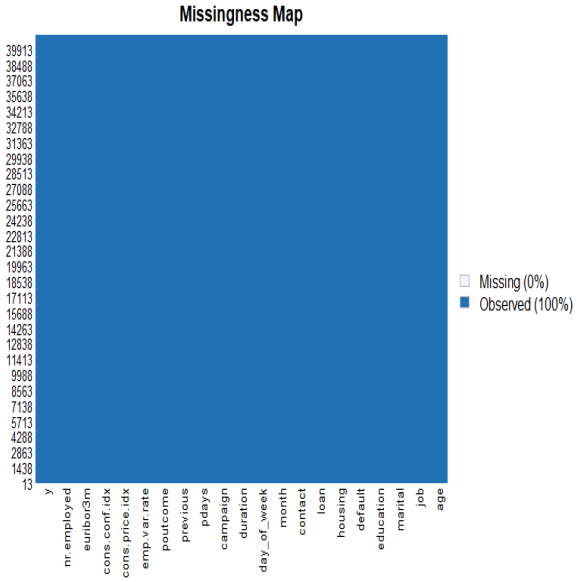

Figure 2:Missingness map for telemarketing dataset

Then the categorical variables are converted to factors and are cleaned. The duration column had extreme outliers and were removed using filtering as there were very less rows.

The predictor variables are checked for multicollinearity using the correlation plot in Fig.1 and the attributes previous and pdays, euribor3m and nr.employed are highly correlated. So they are excluded from the data

Figure 3: Correlation plot for telemarketing data

The data is explored statistically by plotting graphs between the predictors and response variables and found that when the Consumer price index increases, the employee variance rate also increases whereas it decreases when consumer confidence index increases.

iii. Data Mining: The data was also split into training and testing datasets in 80:20 ratio. The following algorithms were used to fit the data and do prediction.
Logistic Regression: Logistic regression model is used when the value to be predicted is binomial. It calculates the probability of the class or in other words likelihood and predict the class. The training data was uesd to train the model and the model is shown in Fig.3.

```
Call:
glm(formula = y ~ ., family = binomial, data = sample.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7324  -0.3023  -0.1883  -0.1355   3.4032
```

Figure 4:Logistic regression model

Support Vector Machines: SVMs tries to find a hyperplane which helps to classify our data into two classes.

```
Call:
svm(formula = as.factor(sample.train$y) ~ ., data = sample.train[, -21], probability = TRUE, kernel = "linear")


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

Number of Support Vectors:  6434

 ( 3226 3208 )


Number of Classes:  2

Levels:
 no yes
```

Figure 5: SVM model

The number of support vectors in the model are 6434 in total. These support vectors are responsible for hyperplanes orientation and the position. The model predicts the probability for each class based on which the labels are chosen.

```
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  6954  179
       yes  615  291

               Accuracy : 0.9012
                 95% CI : (0.8945, 0.9077)
    No Information Rate : 0.9415
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3748

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9187
            Specificity : 0.6191
         Pos Pred Value : 0.9749
         Neg Pred Value : 0.3212
             Prevalence : 0.9415
         Detection Rate : 0.8650
   Detection Prevalence : 0.8873
      Balanced Accuracy : 0.7689

       'Positive' Class : no
```

Figure 6:Confusion matrix for SVM

B. *Loan status in Banking domain*

The objective is to predict the loan status of a customer with the help of many influencing factors using classifiers like random forest.

i. Data Selection: The dataset was fetched from kaggle [20] with dimensions 100514 x 19. The id columns were ignored and only 17 columns were selected which formed the target data.

ii. Pre-processing and Transformation: The data consisted of lot of null values. The column 'months.since,last.delinquement' had null values for 53% of total rows and so it was removed.

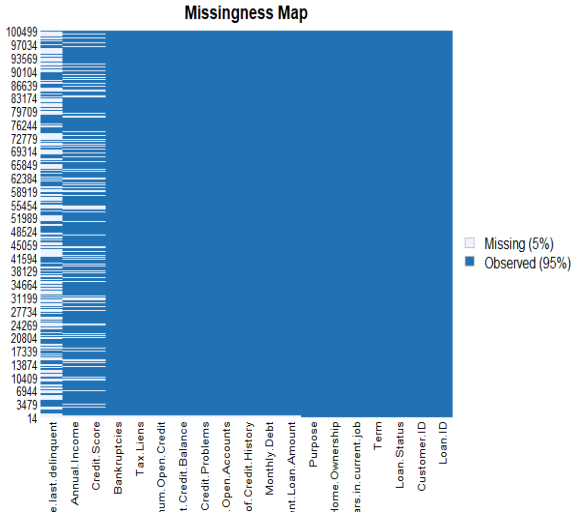

Figure 7 : Missingness map for bank loan dataset

All other columns had relatively less null values that can be removed. The extreme outliers in the columns credit score, current loan amount, credit balance and others were removed. The levels of the factor variables were changed according to R naming rules to run using caret package.

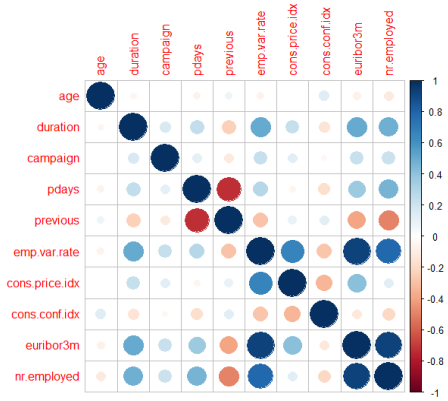

Figure 8: Correlation plot for bank loan data

iii. Data Mining: The training and test data were split in the ratio of 80:20 and the model was built. As our dataset was imbalanced, we used SMOTE sampling technique as in [14]. The random forest model was trained by applying cross validation with k=10 folds and the best mtry value (which is the number of values selected randomly at each split) based on ROC value. By tuning the model the best mtry value was found to be 27 and the model started learning with mtry=27.

```
Random Forest

32950 samples
   20 predictor
    2 classes: 'no', 'yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 29654, 29654, 29655, 29655, 29656, 29655, ...
Addtional sampling using SMOTE

Resampling results across tuning parameters:

  mtry  ROC        Sens       Spec
   2    0.9236472  0.9651138  0.4089347
  27    0.9443878  0.9168201  0.7747942
  53    0.9412962  0.9176754  0.7669739

ROC was used to select the optimal model using the largest value.
The final value used for the model was mtry = 27.
>
```

Figure 9:Random Forest model

iv. Evaluation: The classification models were evaluated by plotting ROC curves, their accuracy values, AUC and the best model was chosen.

C. Airbnb price in Hospitality domain

Regression algorithms like Linear and XGBoost regression were applied to the Airbnb dataset to train the model to predict the price of the room based on various factors at a specific time.

i. Data Selection: The dataset is fetched from [21] which consist of the data about the Airbnb hosts and their listings along with other factors that help us predict the price of the property. The data is of dimension 48895x16 and the data about the location, the neighbourhood, , type of the room , availability= and reviews count are selected and target data is formed. Id columns and host name columns were ignored in the first step.

ii. Data Cleaning and Transformation: The reviews per month column had more missing values.

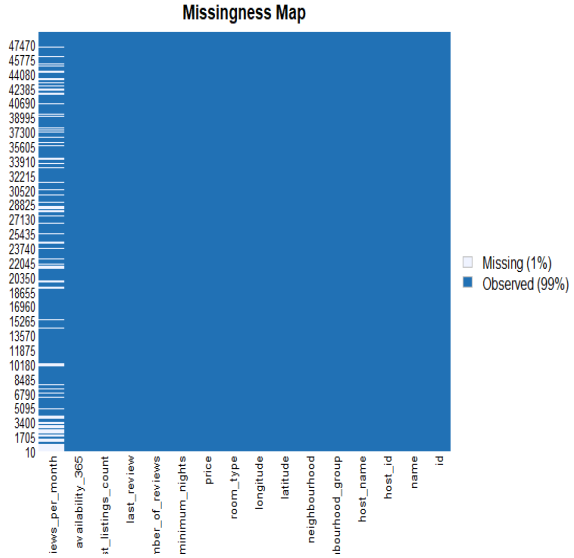

Figure 10: Missingness map of airbnb dataset

The missing values were handled by removing those and the outliers were removed in the price column by using filters as there were very few values.

iii. Data Mining: The final data was divided in such a way that there are 80% of data for training and remaining 20% for testing set.

Linear Regression:

Before applying the linear regression, the assumptions are checked and ensured that it is homoscedastic, no multicollinearity among the predictors and the residuals are normally distributed.

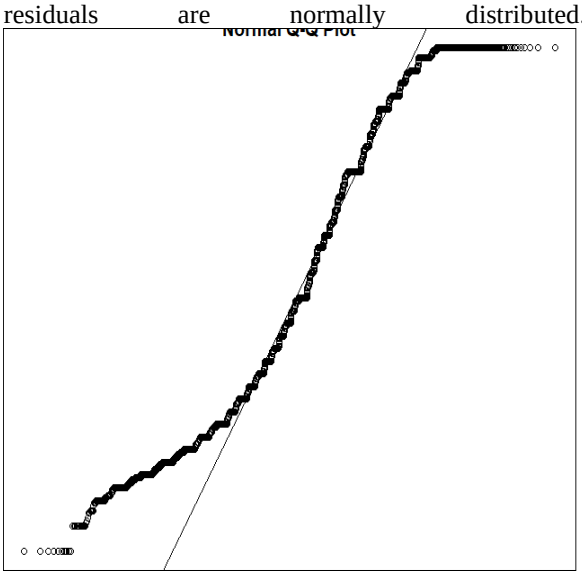

Figure 11: Normal Q-Q plot

The response variable is normally distributed which is important in linear regression.

```
Call:
lm(formula = price ~ ., data = airbnb.train)

Residuals:
     Min       1Q   Median       3Q      Max
-141.496  -18.794   -3.471   16.362  172.405

Coefficients: (4 not defined because of singularities)
                                        Estimate Std. Error  t value Pr(>|t|)
(Intercept)                           -1.855e+04  2.263e+03   -8.193 2.66e-16 ***
neighbourhood_groupBrooklyn           -3.851e+01  9.690e+00   -3.974 7.08e-05 ***
neighbourhood_groupManhattan           3.606e+00  7.907e+00    0.456 0.648410
neighbourhood_groupQueens             -1.623e+01  7.189e+00   -2.258 0.023957 *
neighbourhood_groupStaten Island      -1.139e+02  2.371e+01   -4.805 1.56e-06 ***
neighbourhoodArden Heights            -4.196e+01  2.669e+01   -1.572 0.115956
neighbourhoodArrochar                  1.964e+01  2.208e+01    0.890 0.373668
neighbourhoodArverne                   2.540e+01  7.801e+00    3.256 0.001130 **
neighbourhoodAstoria                   1.845e+00  2.939e+00    0.628 0.530137
neighbourhoodBath Beach               -2.497e+01  9.519e+00   -2.623 0.008724 **
neighbourhoodBattery Park City        -2.372e+01  8.062e+00   -2.942 0.003259 **
neighbourhoodBay Ridge                -2.587e+01  4.829e+00   -5.358 8.50e-08 ***
neighbourhoodBay Terrace               8.265e+01  1.512e+01    5.466 4.64e-08 ***
neighbourhoodBay Terrace, Staten Island 3.065e+00 2.921e+01    0.105 0.916442
neighbourhoodBaychester               -3.610e+00  1.555e+01   -0.232 0.816422
```

Figure 12:Linear Regression model in R

XGBoost Regression:

The XGBoost regression is decision tree based and is gradient boosted. The model was trained with training data and cross validation with 5 folds was done to find the parameters for best model.

```
> xgb_caret$bestTune
   nrounds max_depth  eta gamma colsample_bytree min_child_weight subsample
24    1000         6 0.01     0              1                4         1
```

Figure 13: Best tuning parameters

The best iteration or the value for nround is chosen by doing the cross validation.

```
> xgbcv <- xgb.cv( params = default_param, data = dtrain, nrounds = 2000, nfold = 5,
+                  showsd = T, stratified = T, print_every_n = 40, early_stopping_rounds = 10, maximize = F)
[1]     train-rmse:106.382782+0.059984  test-rmse:106.384103+0.244494
Multiple eval metrics are present. Will use test_rmse for early stopping.
Will train until test_rmse hasn't improved in 10 rounds.

[41]    train-rmse:74.353891+0.032086   test-rmse:74.448817+0.291529
[81]    train-rmse:54.028086+0.015303   test-rmse:54.260967+0.300281
[121]   train-rmse:41.675862+0.020596   test-rmse:42.089524+0.272907
[161]   train-rmse:34.596281+0.030228   test-rmse:35.212927+0.219765
[201]   train-rmse:30.769978+0.037330   test-rmse:31.584682+0.156517
[241]   train-rmse:28.760409+0.040583   test-rmse:29.758443+0.112332
[281]   train-rmse:27.697925+0.042941   test-rmse:28.860546+0.101536
[321]   train-rmse:27.099314+0.039825   test-rmse:28.411391+0.111636
[361]   train-rmse:26.728345+0.033960   test-rmse:28.177953+0.128867
[401]   train-rmse:26.463858+0.033824   test-rmse:28.039815+0.138150
[441]   train-rmse:26.261628+0.033312   test-rmse:27.948511+0.146410
[481]   train-rmse:26.076939+0.028576   test-rmse:27.876246+0.156214
[521]   train-rmse:25.921019+0.024702   test-rmse:27.828109+0.161842
[561]   train-rmse:25.789342+0.022481   test-rmse:27.794336+0.163776
[601]   train-rmse:25.673067+0.022640   test-rmse:27.767387+0.165425
[641]   train-rmse:25.570196+0.021125   test-rmse:27.746982+0.168747
[681]   train-rmse:25.476726+0.018232   test-rmse:27.731397+0.170412
[721]   train-rmse:25.393426+0.017060   test-rmse:27.717646+0.169977
[761]   train-rmse:25.309671+0.018507   test-rmse:27.706402+0.171699
[801]   train-rmse:25.231085+0.020480   test-rmse:27.695680+0.173030
[841]   train-rmse:25.154793+0.020042   test-rmse:27.684561+0.172747
[881]   train-rmse:25.076015+0.026928   test-rmse:27.676718+0.172894
[921]   train-rmse:24.999901+0.034966   test-rmse:27.670321+0.174503
[961]   train-rmse:24.927021+0.042093   test-rmse:27.665107+0.174970
[1001]  train-rmse:24.855394+0.046220   test-rmse:27.660452+0.174938
[1041]  train-rmse:24.785503+0.047526   test-rmse:27.657026+0.175982
Stopping. Best iteration:
[1034]  train-rmse:24.797663+0.045838   test-rmse:27.656807+0.175596
```

Figure 14: Cross validation to find best iteration

The model is trained again with the parameters obtained from tuning.

```
> xgb_mod
##### xgb.Booster
raw: 5.8 Mb
call:
  xgb.train(params = default_param, data = dtrain, nrounds = 1034)
params (as set within xgb.train):
  objective = "reg:squarederror", booster = "gbtree", eta = "0.01", max_depth = "6", gamma = "0", min_child_weight = "4", su
bsample = "1", colsample_bytree = "1", validate_parameters = "TRUE"
xgb.attributes:
  niter
callbacks:
  cb.print.evaluation(period = print_every_n)
# of features: 11
niter: 1034
nfeatures : 11
```

Figure 15: XGB model

IV. Evaluation Methods

The model evaluation is the most important part of model building. Evaluating a model helps us to improve the model iteratively by changing the hyperparameters and tuning the model to increase the performance and also check how tout model will work for a unknown data. I used the test data that was split from every datasets to evaluate the respective model performance.

The classification algorithms used in our research are evaluated using metrics like Area under the curve (AUC), confusion matrix and the accuracy, whereas the regression models was assessed by MSE, R squared, RMSE and MAE.

A. Telemarketing dataset

To evaluate the logistic regression and SVM model the following metrics were used.

Confusion Matrix: The confusion matrix shows the predicted values in a cross table format.. It shows false positives, true positives , false negatives and true negatives.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6939  536
         1  194  370
```

Figure 16- Confusion matrix for Logistic Regression

```
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  6954  179
       yes  615  291
```

Figure 17-Confusion matrix for SVM

The table-1 shows the accuracy, specificity values. The accuracy has been chosen as it provides how accurate the classification is. Accuracy along with sensitivity and specificity gives a better understanding to know about the false negatives and positives.

Metric	Logistic Regression	SVM
Accuracy	0.9092	0.9012
Sensitivity	0.9728	0.9187
Specificity	0.4084	0.6191
AUC	0.7922	0.7689

Table 1:Evaluation metrics comparison for models on telemarketing dataset

The ROC curve can be used to visualize how well the classifier performance for each classification threshold. AUC can be calculated and is the area under the ROC curve.

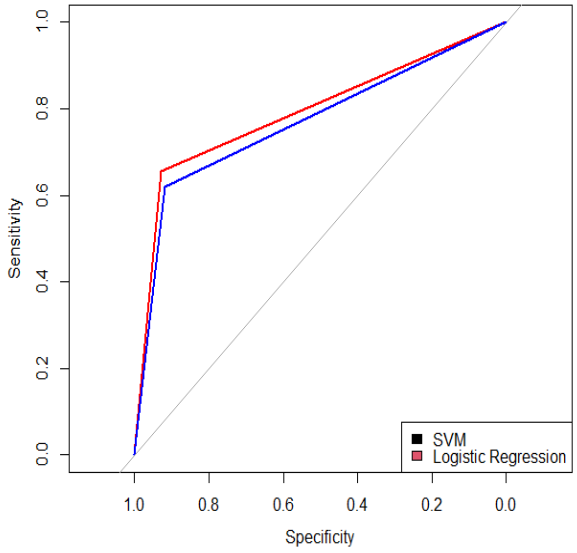

Figure 18: ROC curves for SVM and Logistic Regression

From the above evaluations we can see that the logistic regression performed well in all aspects with an accuracy of 90.92 %. The AUC is 79.22 % which means that our model has higher chance of predicting the classes correctly.

B. Loan Status Prediction

The Random forest classifier is trained to predict the loan status of the customers and was validated using test data. The following metrics were used to evaluate the model.

The Confusion matrix of the model provided accuracy, specificity and sensitivity which helped in determining how well the model performs in classifying the datasets.

The SMOTE method is used which is a oversampling technique to tackle the imbalances in the data. It reduces the problem by oversampling the minority class and trains the algorithm. In our dataset, the loan status is imbalanced with 3:1 ratio of positive and negative classes. So we use 10 fold cross validation technique with SMOTE to tune our model and found the best mtry value to be 27. The algorithm is trained with that value and model is built. It is evaluated by predicting the test data and looking at the evaluation metrics below:

```
Confusion Matrix and Statistics

          Reference
Prediction    Charged.Off Fully.Paid
  Charged.Off         822       1800
  Fully.Paid         1377       7373

               Accuracy : 0.7206
                 95% CI : (0.7123, 0.7289)
    No Information Rate : 0.8066
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1655

 Mcnemar's Test P-Value : 7.05e-14

            Sensitivity : 0.37381
            Specificity : 0.80377
         Pos Pred Value : 0.31350
         Neg Pred Value : 0.84263
             Prevalence : 0.19337
         Detection Rate : 0.07228
   Detection Prevalence : 0.23057
      Balanced Accuracy : 0.58879

       'Positive' Class : Charged.Off
```

Figure 19:Random Forest Model

The specificity value is high but the sensitivity is not so high. Even though the accuracy is satisfactory, the number of False positives are more. As this is not a health related domain the risk is less. The customers who get classified as false positive may get the loan but still there is no surity whether they will pay without default. Customers who might be repaying the loan back might not get it as they are falsely classified.

C. Airbnb Price Prediction

The Price prediction of Airbnb data involved implementing regression algorithms like Linear Regression and XGBoost Regression and they were evaluated with the help of the test set and with following metrics:

Mean Square Error (MSE) : It is a measure which shows how close our data points are to the fitted regression line

Root Mean Square Error (RMSE) : It is the measure of how much the residuals are spread out.

Mean Absoute Error (MAE) : It is the mean of the difference of observed value and the actual value.

These three metrics are compared along with the R squared value to compare the best algorithm.

In our study, the XGBoost algorithm

Metrics	Linear Regression	XGBoost Regression
MSE	955.27	763.12
MAE	24.28	21.31
RMSE	30.9	27.62

Table 2: Evaluation metrics for Regression

From the above values, we can see that the error rate for XGBoost is better compared to the linear regression.

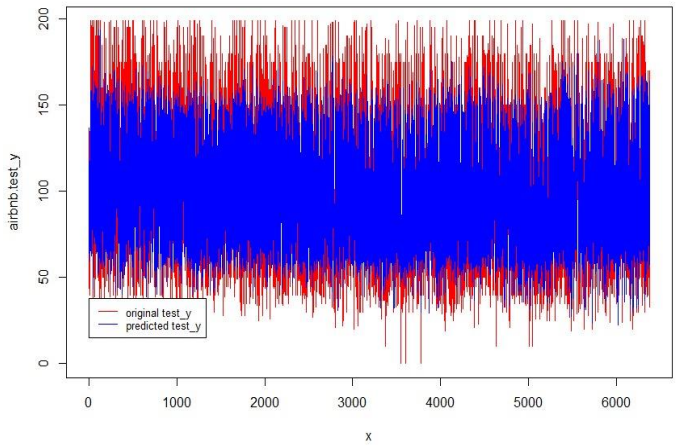

Figure 20: Predicted vs Actual price

The RMSE of 27.62 means that the predicted value differs with a average error of 27.62 which is very good margin to predict the hotel prices.

V. Conclusion

The paper implemented classification and regression methodologies to three different datasets and trained and evaluated the models based on evaluation criteria like ROC, AUC, MSE and MAE. The KDD methodology was followed from the domain understanding to the predictions made using the models built. In the telemarketing problem, the logistic regression performed better with an accuracy of 90.92% in comparison with SVM. The classification algorithms were tuned using 10 fold cross validation, by which random forest achieved an accuracy of 72% in predicting loan status of customer. In regression problem of predicting Airbnb price, both linear regression and XGBoost were used. The parameters of XGBoost model was tuned to achieve the best model.

References

[1] S. Moro, P. Cortez and P. Rita, "A data-driven approach to predict the success of bank telemarketing", *Decision Support Systems*, vol. 62, pp. 22-31, 2014.

[2] E. Zeinulla, K. Bekbayeva and A. Yazici, "Comparative study of the classification models for prediction of bank telemarketing," 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), Almaty, Kazakhstan, 2018, pp. 1-5, doi: 10.1109/ICAICT.2018.8747086.I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] G. Arutjothi and C. Senthamarai, "Prediction of loan status in commercial bank using machine learning classifier," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, 2017, pp. 416-419, doi: 10.1109/ISS1.2017.8389442.

[4] Z. Peng, Q. Huang and Y. Han, "Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm," 2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT), Jinan, China, 2019, pp. 168-172, doi: 10.1109/ICAIT.2019.8935894.

[5] D. Devi, S. K. Biswas and B. Purkayastha, "A Cost-sensitive weighted Random Forest Technique for Credit Card Fraud Detection," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-6, doi: 10.1109/ICCCNT45670.2019.8944885.

[6] C. R. Madhuri, G. Anuradha and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," 2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICSSS.2019.8882834.M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[7] Y. Li, "Credit Risk Prediction Based on Machine Learning Methods," *2019 14th International Conference on Computer Science & Education (ICCSE)*, Toronto, ON, Canada, 2019, pp. 1011-1013, doi: 10.1109/ICCSE.2019.8845444.

[8] Krichene, Aida. (2017). Using a naive Bayesian classifier methodology for loan risk assessment: Evidence from a Tunisian commercial bank. Journal of Economics, Finance and Administrative Science. 22. 3-24. 10.1108/JEFAS-02-2017-0039.

[9] D. S. Sisodia, S. Vishwakarma and A. Pujahari, "Evaluation of machine learning models for employee churn prediction," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 1016-1020, doi: 10.1109/ICICI.2017.8365293..

[10] A. S. Galathiya, A. P. Ganatra and C. K. Bhensdadia, "Classification with an improved decision tree algorithm", *Int. J. Comput. Appl.*, vol. 46, no. 23, pp. 1-6, 2012.

[11] J. Kim et al., "Optimal Feature Selection for Pedestrian Detection Based on Logistic Regression Analysis," 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, 2013, pp. 239-242, doi: 10.1109/SMC.2013.47.

[12] M. Sivasakthi, "Classification and prediction based data mining algorithms to predict students' introductory programming performance," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 346-350, doi: 10.1109/ICICI.2017.8365371.

[13] Kavitha S, Varuna S and Ramya R, "A comparative analysis on linear regression and support vector regression," 2016 Online International Conference on Green Engineering and Technologies (IC-GET), Coimbatore, 2016, pp. 1-5, doi: 10.1109/GET.2016.7916627.

[14] T. Lu, Y. Huang, W. Zhao and J. Zhang, "The Metering Automation System based Intrusion Detection Using Random Forest Classifier with SMOTE+ENN," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, China, 2019, pp. 370-374, doi: 10.1109/ICCSNT47585.2019.8962430.

[15] A. El-Koka, K. Cha and D. Kang, "Regularization parameter tuning optimization approach in logistic regression," 2013 15th International Conference on Advanced Communications Technology (ICACT), PyeongChang, 2013, pp. 13-18.

[16] Tom Fawcett, An introduction to ROC analysis, Pattern Recognition Letters, Volume 27, Issue 8, 2006, Pages 861-874, ISSN 0167-8655

[17] Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge University Press. doi:10.1017/CBO9781107298019

[18] M. C. M. Oo and T. Thein, "Hyperparameters Optimization in Scalable Random Forest for Big Data Analytics," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019, pp. 125-129, doi: 10.1109/CCOMS.2019.8821752.

[19] http://archive.ics.uci.edu/ml/datasets/Bank+Marketing# [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

[20] https://www.kaggle.com/zaurbegiev/my-dataset

[21] http://insideairbnb.com/get-the-data.html