Applications of Data Mining and Machine Learning Models

Abstract— **Machine Learning (ML) models have been gaining importance amongst researchers over the past couple of years and has become an area which is continuously increasing its reach and has found its application in most branches of engineering and science. In the present work, I have looked into two major classes of ML models, (a) Classification Model and (b) Regression model. In this work, I have explored four different types of Classification Model, (a) Decision Tree, (b) Extreme Gradient Boost (XGBoost), (c) Random Forest and (d) Naïve-Bayes Model; and one type of Regression model, i.e. Multiple Linear Regression Model. For the purpose of classification model, I have considered two different datasets. One of the datasets consists of fraudulent online transaction and the other dataset comprises of passenger satisfaction survey for a United Stated airline. For regression, I used a diamond dataset that comprises of a number of properties of a diamond and its associated price. I have developed a prediction model that predicts the price of a diamond based on its properties. In all the cases, the model is trained on a chosen sample data and the test set is used for measuring model accuracy.**

Keywords—Random forest, Decision tree, XGBoost, Naïve Bayes, Regression

I. Introduction

Machine Learning (ML) is an area of computational science which has its roots as early as 1943 [1]. However, with the huge advancement in computational technology both in terms of computational speed and memory over the last few decades, which includes the use of multi-core CPUs or interconnected clusters and more recently GPUs with several thousand cores using parallel architecture, ML has gained much more prominence. Much research has already been carried out in this area and researchers are working further to improve upon the currently existing models. As such, the present work focusses on the application of ML. In the subsequent paragraphs, a brief description of this work highlighting the research motivation has been presented.

With the onset of the digital age and everything going online, financial frauds have become more and more common. A recent report by PwC states that "51% of surveyed organisations say they experienced fraud in the past two years, the highest level in our 20 years of research" [2]. Of this, 25% are reported to have lost over $1 million, a data which is of immense concern. This aspect has motivated the first part of my work where a ML model has been developed to predict whether a certain transaction is fraudulent or not.

The next important aspect that has been looked at is the application of ML to business. A business cannot run successfully if the customer is not satisfied and there are many examples where this can be seen. In the present work, I have therefore chosen customer satisfaction in the airlines business. With an increase in the number of competitors, it is essential that airlines are able to predict the satisfaction level of their customers. It might so happen that a customer is satisfied with the service provided by the airline if they simply got complimentary snacks for short journey flights. On the other hand, there might be certain customers who are dissatisfied simply because airlines did not provide them with internet connection for similar flights. As can be seen the feedback of the passenger is highly dependent on the facilitates the passenger is expecting from the airlines which is not something that entirely depends on customer's priority or taste. In my work, I have tried to explore the application of ML such that the model can predict the feedback that a customer will probably give and whether the airlines should act on that to improve that particular aspect of its service.

Diamonds are one of the most precious items in the world. However, the quality of the demand depending on its different kind of properties is something that might reduce or increase its value greatly. For a diamond merchant, it is essential that to offer a competitive and fair price to the customers, the merchant should be able to predict the price of the diamond correctly. Hence, in the final part of my work, I have undertaken the study of price prediction for diamonds depending on its properties.

Having realized the broad scope of the work, this report begins with a brief literature review of the related works in Section II. Subsequently, in Section III, a few data mining methodology is covered followed by conclusions and future directions in Section IV.

II. Related Work

In our daily life, we often face situations where we need to take decisions based on historical data, be it predicting the stock market analyzing the past market records or identifying the type of sentiment a social media customer is likely to show towards a certain article or post. With the increasing usage of technologies day by day, analyzing the massive amount of data has been a real challenge over the past few decades. However, with the advent of ML models in the form of classifiers and regressors this decision making tasks has been proved to be easier.

Decision tree is one of the easiest ways to maintain an organized database. It has been proved to be efficient in terms of business management [3], customer relationship management [4], detecting fraud financial statements to enhance Government's tax income [4]. Another very important real life scenario where decision trees play a vital role is the engineering sector. [3] shows how decision tree based approaches help estimating the amount of energy is likely to be consumed by the individuals. Application of decision trees can also be seen in faulty machinery detection

[5], improving the management of healthcare services [3] [5] [4]. The wide variety of successful application of decision trees motivated me to choose this technique as one of my classifier in this work.

With its own capabilities, XGBoost is a weighted quantile sketch algorithm that uses the gradient boosted decision tree as the base for its overall implementation. In addition to all the problems that can be solved by decision trees, XGBoost shows its merit in classifying images [6], validating faces [7]. A wide range of classification problems has been solved with the help of random forest and naïve bayes, such as, detecting credit card frauds, segmenting customers for a better estimation of how the company's distribution program can better be configured to suit each client's needs [8]. Not only that, random forest has been proved to be very useful in healthcare and medicine sectors in terms of predicting complex mixture of individual chemicals in a medicine [8] [9] or [10] and [11] shows how it analyzes patient records for a better health care.

Many businesses and their top executives are now adopting regression analysis (and other types of statistical analysis) to make better business decisions and reduce guesswork and gut instinct [12]. This is because regression enables firms to take a scientific approach to management. Both small and large enterprises are frequently bombarded with an excessive amount of data. Managers may use regression analysis to filter through data and choose the relevant factors to make the best decisions possible [13]. All the above mentioned merits of ML classifiers and regressors motivated me to apply them in my work albeit for different datasets with different objectives.

III. Data Mining Methodology

Everyday there is more information generated by business transactions, scientific research, sensor data, pictures, movies, and other things than we can handle. Therefore, in order to make better decisions, we need a system that can automatically generate reports, views, or summaries of data while also extracting the key information from the available information. This is called data mining, also known as knowledge discovery in databases (KDD). Formally speaking, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.

KDD undergoes nine consecutive steps [14], such as, (1) understanding the dataset, (2) data selection, (3) data cleaning and pre-processing, (4) transforming the data into desired form, (5) selecting the appropriate data mining task, (6) right choice of the data mining algorithm, (7) then application of the selected data mining algorithm, (8) model evaluation and lastly, (10) interpreting the outcome of the model.
Next, we discuss the datasets used to conduct experiments in this project and the model's performance individually.

A. Dataset 1

In recent days, *credit card fraud* is a very common term. By credit card fraud we mean that someone has used another person's credit card or account information to make unauthorized purchases or access funds. The first dataset of the study contains historical information on fraudulent transactions at a certain period of time which can be found on *kaggle*.

1) Data Description - This data contains a total of 1,15,34,325 data points distributed in 11 different attributes and 10,48,575 rows. The abbreviated attribute names and their corresponding actual meanings have been listed below

TABLE I. VARIABLE DESCRIPTION (DATASET 1)

Variable Name	Actual Meaning
step	A unit of time where 1 step equals 1 hour
type	Type of online transaction
amount	Amount of the transaction
nameOrig	Origin of the transaction
oldbalanceOrg	Balance before transaction at origin
newbalanceOrg	Balance after transaction at origin
nameDest	Destination of transaction
oldbalanceDest	Initial balance before transaction at destination
newbalanceDest	New balance after transaction at destination
isFraud	'0' – not fraud transaction '1' – fraud transaction

2) Objective - My main objective is to train two machine learning models for classifying fraudulent and non-fraudulent transactions based on this dataset and testing the models' performance and accuracy. Finally, compare the models and conclude the best classification model.

3) Data Pre-processing - *Firstly,* the whole dataset has been filtered and only the variables of interest (eg. type, amount, oldbalanceOrg, newbalanceOrg, oldbalanceDest, newbalanceDest, isFraud etc.) have been stored after renaming them for a better understanding.

Secondly, I summarized the filtered dataset and checked for any white spaces or missing values. But none of them was found.

Finally, I tried to find out the proportion of fraudulent and non-fraudulent transactions in the dataset and noticed that nearly 99.89% of transactions were non-fraudulent and nearly 0.11% (less than 1%) of the total transactions belong to fraudulent transactions. So this dataset is highly imbalanced and hence I used *Stratified Sampling* to maintain the population proportions of binary responses in the sample data and split 75% of the data into the training set and 35% of the data into the testing set. But still, I found that the imbalance was maintained. So I have decided to *downsample* the training and testing dataset i.e. I have randomly chosen 2,000 non-fraudulent data among 7,85,561 and 666 among 2,61,872 from the training and testing set respectively to balance the proportion of responses in the target variable.

4) Model 1 (Decision Tree) - I have fitted a Decision Tree model on the training dataset for binary classification using *rpart* package. Here ‘isFraud’ is the target variable i.e. the variable I want to classify based on the other independent variables.

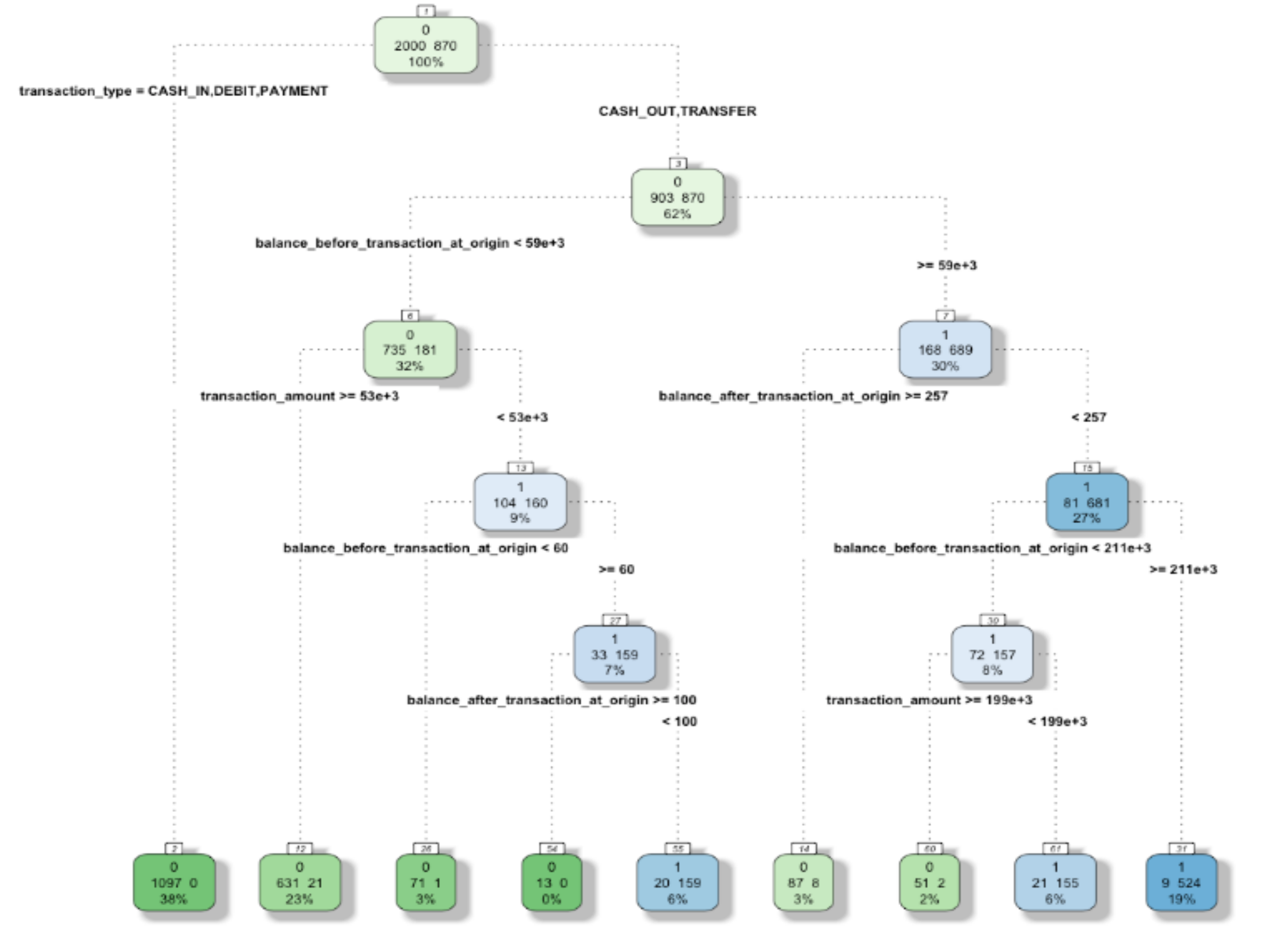

1
0
2000 870
100%
transaction_type = CASH_IN,DEBIT,PAYMENT
CASH_OUT,TRANSFER
3
0
903 870
62%
balance_before_transaction_at_origin < 59e+3
>= 59e+3
6
0
735 181
32%
7
1
168 689
30%
transaction_amount >= 53e+3
< 53e+3
balance_after_transaction_at_origin >= 257
< 257
13
1
104 160
9%
15
1
81 681
27%
balance_before_transaction_at_origin < 60
>= 60
balance_before_transaction_at_origin < 211e+3
>= 211e+3
27
1
33 159
7%
30
1
72 157
8%
balance_after_transaction_at_origin >= 100
< 100
transaction_amount >= 199e+3
< 199e+3
2
0
1097 0
38%
12
0
631 21
23%
26
0
71 1
3%
54
0
13 0
0%
55
1
20 159
6%
14
0
87 8
3%
60
0
51 2
2%
61
1
21 155
6%
31
1
9 524
19%

Fig. 1. Decision Tree

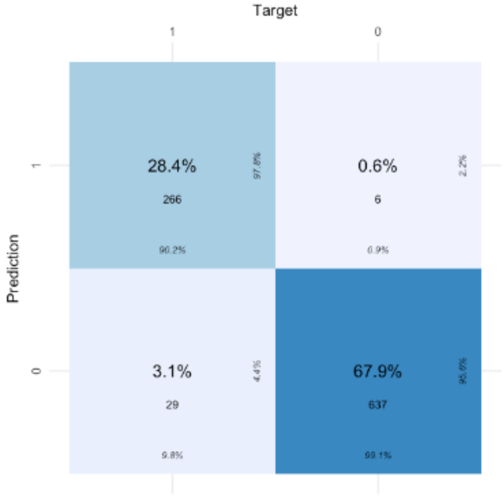


Fig. 2. Confusion Matrix of Decision Tree

4.1) Model Summary: The model has used a total of 8 splits to get the final optimum result. Starting from the root node, each split has been done in such a way that the *Information Gained* can be maximized (alternatively, entropy can be minimized). At the root node (from Fig.1), the total number of non-fraudulent transactions was 2000 and 870 transactions were fraudulent transactions. Now, by maximizing the information gained, the first split has been done using the ‘transaction_type’ variable. If the transaction types were ‘cash_in’, ‘debit’ or ‘payment’ then 1097 transactions can be classified as non-fraudulent and have no fraudulent transactions at all which leads to the leaf node. But if the transaction types were ‘cash_out’ or ‘transfer’, then we have a set of 903 non-fraudulent and 870 fraudulent transactions. Now the next split has been done on the basis of the variable ‘balance_before_transaction_at_origin’ so that the entropy could be minimized and in a similar fashion, we can proceed to the next split until we get the optimum tree.

4.2) Model Performance and Accuracy: When I tried to feed this model with the testing dataset, it classified 637 non-fraudulent transactions and 266 fraudulent transactions correctly whereas only a total of 35 transactions were missclassified which leads to the model accuracy of 96% (approx). For this model, Cohen's Kappa is 0.91 which means that our classifier has 91% better performance than other classifiers that simply guess at random according to the frequency of each class. Fig. 2 clearly shows that 67.9% of non-fraudulent transactions and 28.4% of fraudulent transactions in the test data are correctly classified by the model.

4.3) ROC Curve and AUC: A *receiver operating characteristic (ROC)* curve is a graph showing the performance of the classification model at all classification thresholds. It is basically the graphical representation of the trade-off between the false negative and false positive rates at every possible cut-off point. The plot shows the false positive rate i.e. (1-specificity) on the x-axis and the true positive rate i.e. sensitivity on the y-axis. Here, for this model, the cut-off at the yellow region (in Fig. 3) clearly signifies the optimal decision threshold since the false positive rate is minimum whereas the true positive rate is maximum at that point and the *Area Under the ROC Curve (AUC)* is 0.98 (approx.)

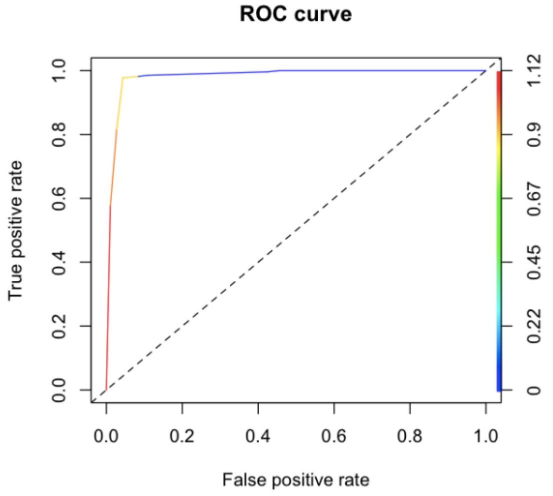


Fig. 3. ROC curve for Decision Tree

5) Model 2 (XGBoost) – I have fitted a second model, namely Extreme Gradient Boosting (XGBoost) Model for binary classification. Though XGBoost also uses Decision Tree as base learners but the trees used by the XGBoost are called CART trees (Classification and Regression trees) and instead of containing a single decision tree in each leaf node, they contain real value scores of whether an instance belongs to a group. After the tree reaches maximum depth, the decision is made by converting the scores into categories using a certain threshold.

5.1) Model Summary: I have tried to fit the XGBoost model using the 'caret' package in R (*while modelling, I have noted that the internal algorithm of this package tries to convert the factor levels to the predefined levels in the algorithm. For instance, we had two factors in our target variable ('inFraud'), namely '0' for non-fraudulent transactions and '1' for fraudulent transactions. Now with these numeric factor levels, it could not fit the model since the algorithm failed to convert that to the predefined factor levels. So I had to change the factor levels as per the defined convention in the package*). The model has been trained through 5-fold cross-validation where in each iteration, 2,870 samples are used with 6 predictor variables. Here a set of tunning parameters has been used but the model with nrounds (number of trees) = 1500, max_depth (depth for each tree) = 6, eta (learning rate) = 0.3, gamma (pruning parameter) = 1, colsample_bytree (subsample ratio of columns for tree) = 1, min_child_weight = 1 and subsample = 1 has shown the best performance with 98% accuracy.

5.2) *Model Performance and Accuracy*: The model has shown great performance in classifying fraudulent and non-fraudulent transactions for the testing dataset. It classified 658 non-fraudulent transactions and 268 fraudulent transactions correctly whereas only a total of 12 transactions were miss classified which leads to the model accuracy of 98% (approx). For this model, Cohen's Kappa is 0.96 which means that our classifier has 96% better performance than other classifiers that simply guess at random according to the frequency of each class. Fig. 4 clearly shows that 70.1% of non-fraudulent transactions and 28.6% of fraudulent transactions in the test data are correctly classified by the model.

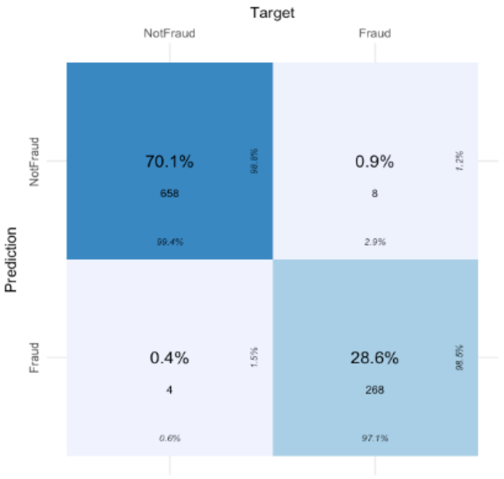


Fig. 4. Confusion Matrix of XGBoost

5.3) ROC Curve and AUC: For this model, the cut-off at the green region (in Fig. 5) clearly signifies the optimal decision threshold since the false positive rate is minimum whereas the true positive rate is maximum at that point and the Area Under the ROC Curve (AUC) is 0.99 (approx.).

6) Evaluation and Conclusion: As discussed above, in Dataset 1, I have fitted two classification models, namely Decision Tree and XGBoost to classify fraudulent transactions. Both our models have performed greatly and shown considerably large accuracy values but judging by the models' accuracy values, the ROC curve and the area under the ROC curve, *it seems that XGBoost has done a better job in classifying fraudulent transactions for this dataset.*

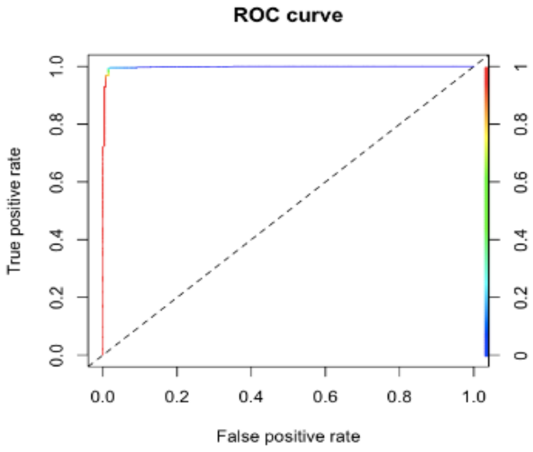


Fig. 5. ROC curve for XGBoost

B. Dataset 2

My second dataset is about the *USA Airline Passenger* satisfaction survey which contains a total of 25,97,600 data points distributed in 25 different attributes and 1,03,904 rows. This data has been selected from *kaggle*, an open data source.

1) Data Description: This dataset contains 1,03,904 customer reviews on whether they are satisfied or dissatisfied with the overall airline service based on different parameters. It has *Nominal Data* (eg. gender, customer type, type of travel etc.), *Ordinal Data* (eg. ease of online booking, gate location, legroom service etc.) and *Numerical Continuous Data* (eg. age, flight distance, departure delay in minutes etc.). Each ordinal data is in the form of different scores, starting from 1 (low) to 5 (high) where 0 stands for Not Applicable. The abbreviated attribute names and their corresponding actual meanings have been listed in the following table.

TABLE II. VARIABLE DESCRIPTION (DATASET 2)

Variable Name	**Actual Meaning**
Id	Unique identification number
Gender	Gender of the passengers (Female, Male)
Customer Type	Type of the Customer (Loyal, Disloyal)
Age	The actual age of the passengers
Type of Travel	Purpose of the flight of the passengers (Personal Travel, Business Travel)
Class	*Travel class in the flight of the passengers (Business, Eco, Eco plus)*
Flight Distance	Total distance of the journey
Inflight Wifi Service	Satisfaction level with the inflight wifi service (0-5)
Departure/Arrival time convenient	Satisfaction level with Departure/Arrival time convenient (0-5)
Ease of Online Booking	Satisfaction level with online booking (0-5)
Gate Location	Statisfaction level with gate location (0-5)
Food and Drink	Satisfaction level with food and drink (0-5)
Online Boarding	Satisfaction level with online boarding (0-5)
Seat Comfort	Satisfaction level with seat comfort (0-5)
Inflight Entertainment	Satisfaction level with inflight entertainment (0-5)
On-Board Service	Satisfaction level with on-board service (0-5)
Check-in Service	Satisfaction level with check-in Service (0-5)
Inflight Service	Satisfaction level with inflight service (0-5)
Leg Room Service	Satisfaction level with legroom service (0-5)
Baggage Handling	Satisfaction level with baggage handling (0-5)
Cleanliness	Satisfaction level of cleanliness (0-5)
Departure Delay in Minutes	Minutes delayed while departure
Arrival Delay in Minutes	Minutes delayed while arrival
Satisfaction	Airline Satisfaction Level (Satisfied, Nutral/Dissatisfied)

2) Objective - My main objective is to train two machine learning models for classifying passengers' satisfaction as 'Satisfied' and 'Neutral or Dissatisfied' based on this dataset and testing the models' performance and accuracy. Finally, compare the models and conclude the best classification model.

3) Data Pre-processing - Firstly, the first two columns, namely 'X' (count) and 'id' (unique id of the passengers), have been filtered out from the dataset since those have no significant effect on the target variable ('satisfaction').

Secondly, I have summarised the variables of interest and noticed that all ordinal variables were either in integer type or in character type. So I changed all of them to factors.

Thirdly, I found 310 NA values in the variable 'Arrival.Delay.in.Minutes' and I located the row numbers at which those missing values were placed. Then instead of removing those entire rows, I have *imputed* those missing values with the *median* value of that particular column.

Finally, I tried to find the proportion of 'satisfied' and 'neutral or dissatisfied' responses in the dataset and found that nearly 57% of the total responses were 'neutral or dissatisfied' and nearly 43 % of the total responses were 'satisfied'. Since the dataset was well-balanced, I randomly split the whole set into training and testing sets in an 80-20 ratio.

4) Model 3 (Random Forest) – Random Forest classifier has been used as my first model for this dataset. Though random forest uses decision tree as a base learner, it uses multiple numbers of decision trees to improve the learning outcome of the overall model. Here, the column 'satisfaction' has been used as the target variable to build the model based on other independent variables on the training dataset.

4.1) Model Summary: Firstly, I have built a random forest model using the 'randomForest' package in R to classify the passengers' responses on service satisfaction based on the other attributes in the training dataset. Here a total of 500 decision trees have been used to get the optimum result and 4 variables have been tried in each split for each decision tree. The model correctly classified 46,170 passengers as 'neutral or dissatisfied' and 33,827 passengers as 'satisfied' whereas a total of 3126 responses were misclassified based on the training dataset. *The out-of-bag estimate of the error rate for this model is 3.76%.*

Secondly, I tried to evaluate the *Out Of Bag (OOB)* error rate along with the error rate of misclassifying 'satisfied' and 'neutral/dissatisfied' responses and plot them for each tree (in Fig. 6).

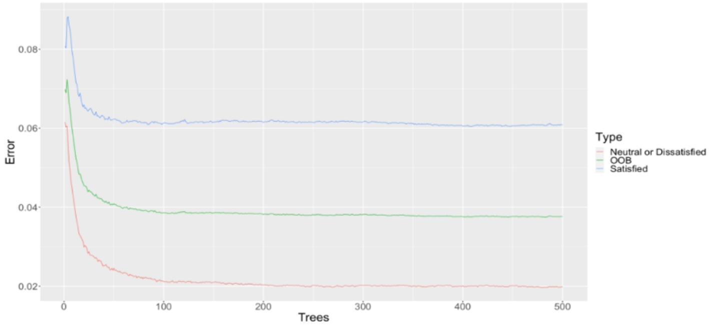


Fig. 6. Number of Trees vs Error rates

Thirdly, I have tried to build another random forest model using a total of 1000 decision trees to verify whether the previous model with 500 trees was optimal or not. Here the out-of-bag estimate of the error rate which is equal to 3.76% has not been improvised and hence it can be concluded that the previous model with 500 decision trees has shown the optimum performance.

Finally, to verify whether 4 variables in each split are optimum or not I tried to iterate the random forest model with 500 trees starting from 1 variable in each split to 10 variables in each split and getting the error rate for each iteration. The following table (Table) suggests that the error rate has significantly dropped up to 4th iteration and after that, it becomes nearly steady. Hence the model with 4 variables in each split has the optimum result.

TABLE III. ERROR RATE

No of Variables in each Split	Error Rate
1	0.06256
2	0.04292
3	0.03915
4	0.03765
5	0.03687
6	0.03613
7	0.03595
8	0.03588
9	0.03550
10	0.03534

4.2) Model Performance and Accuracy: The Random Forest model with a total number of 500 trees has performed quite well on the testing dataset. It classified 11,443 'neutral or dissatisfied' responses and 8371 'satisfied' correctly whereas only a total of 967 responses were miss classified which leads to the model accuracy of 95% (approx). For this model, Cohen's Kappa measure is 0.90 which means that our classifier has 90% better performance than other classifiers that simply guess at random according to the frequency of each class. Fig. 7 clearly shows that 70.1% of non-fraudulent transactions and 28.6% of fraudulent transactions in the test data are correctly classified by the model.

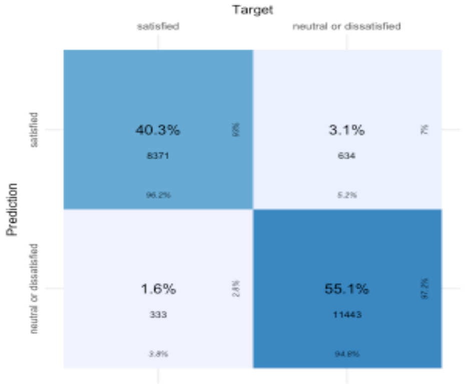


Fig. 7 Confusion Matrix of Random Forest

5) Model 4 (Naive Bayes) – Naive Bayes method is a supervised learning algorithm based on applying *Bayes Theorem* with the 'naive' assumption of conditional independence between every pair of features given the value of the target class variable. In this model, I tried to train the model to predict the passenger satisfaction responses i.e. whether the passengers are satisfied or not given the other attributes present in the train data.

5.1) Model Summary: Here I have used 10-fold cross-validation method to get the optimal outcome. This model has achieved nearly 88% accuracy in predicting the passengers' satisfaction responses on the training dataset using a total number of 83,123 samples and 22 predictors.

5.2) Model Performance and Accuracy: This model has shown a satisfactory performance on the testing dataset with an 89% accuracy level. It has successfully predicted 10,378 responses as 'neutral or dissatisfied' and 8,128 responses as 'satisfied' based on the 22 attributes in the testing dataset. But this model misclassified 1,398 responses as 'satisfied' (the actual response was 'neutral or dissatisfied') and 877 responses as 'neutral or dissatisfied' (the actual response was 'satisfied'). The following figure (Fig. 8) represents the percentage measures of the confusion matrix for this model on testing data.

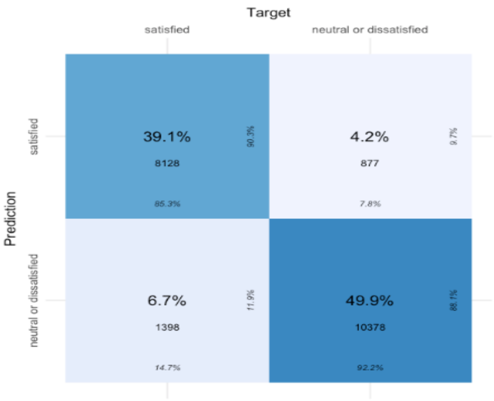


Fig. 8. Confusion Matrix of Naive Bayes

6) Evaluation and Conclusion: As discussed above, Two classification models have been fitted, namely *Random Forest* and *Naive Bayes* to predict the passengers' satisfaction responses. Both the models have performed satisfactory and shown a considerably large accuracy level but judging by the models' accuracy values and the true positive rate and false negative rate based on the test data, *it seems that Random Forest has done a better job in predicting the passengers' response on service satisfaction.*

C. Dataset 3

My third dataset is about the *Price of Diamonds with different measures* which contains a total of 5,93,340 data points distributed in 11 different attributes and 53,940 rows. This data has been selected from *kaggle*, an open data source.

1) Data Description – The dataset contains different physical properties and dimensions of 53,940 diamonds along with their prices. The following table shows the attribute names in the data and their corresponding actual meaning.

TABLE IV. VARIABLE DESCRIPTION (DATASET 3)

Variable Name	**Actual Meaning**
X	Index Counter
Carat	Weight of the diamond in carat
Cut	Cut quality of the diamond. Quality in increasing order Fair, Good, Very Good, Premium, Ideal
Color	Color of the diamond with D being the best and J being the worst
Clarity	How obvious inclusions are within the diamond (in order from best to worst, FL = flawless, I3 = level 3 etc.)
Depth	The height of the diamond measured from the cutlet to the table, divided by its average girdle diameter
Table	The width of the diamond's table expressed as a percentage of its average diameter
Price	Price of the diamond
x	Length in mm (0 to 10.74)
y	Width in mm (0 to 58.9)
z	Depth in mm (0 to 31.8)

2) Objective - Here the prime objective is to train a regression model to predict the price of the diamond based on its physical characteristics using this sample dataset and test the model performance and accuracy.

3) Data Pre-processing - Firstly, the index counter variable has been dropped from the dataset since it has no effect on the target variable 'price'. Then I noticed that some of the factor-type variables were stored in character type (e.g. cut, color, clarity). So the variable types for those three variables have been changed to factors.

Secondly, I renamed the last three columns, namely 'x', 'y' and 'z', to 'length', 'width' and 'depth' for a better understanding. In the descriptive statistic measures for these three variables, some of the observations can be noticed as zero which does not make any logical sense because these three measures define the dimensions of the diamond and if one of them becomes zero that means that the object has now only two dimensions (2-D) and that cannot be possible. So I have dropped those data rows which had at least one of the dimension measure zero.

Thirdly, I have tried to find out the correlation between all the numeric variables, present in the sample data, with the target variable price and picked up those variables which had a high correlation for the final regression model.

Fourthly, I have used the scatter plot to visualize the relationship between our variables of interest and the target variable. A significant number of outliers has been noticed in the boxplot of individual variables which can affect our regression model. For that, I have removed all the values that lie beyond the *[lower 25% - 3*IQR, upper 75% + 3*IQR]* range.

Finally, the processed data has been split into training and testing sets in an 80-20 ratio.

4) Model 5 (Multiple Linear Regression) - Taking 'price' as the response variable, I have built a regression model on the other predictor variables in the training dataset. But here I noted that our response variable contains large values as compared to the other values in the predictor variables and Fig. 9(a) clearly shows that the distribution of the response variable is *right skewed.* So a logarithmic transformation of the response variable has been used to make the distribution symmetric about the mean (Fig. 9(b)) and the regression model has been built with this new response variable.

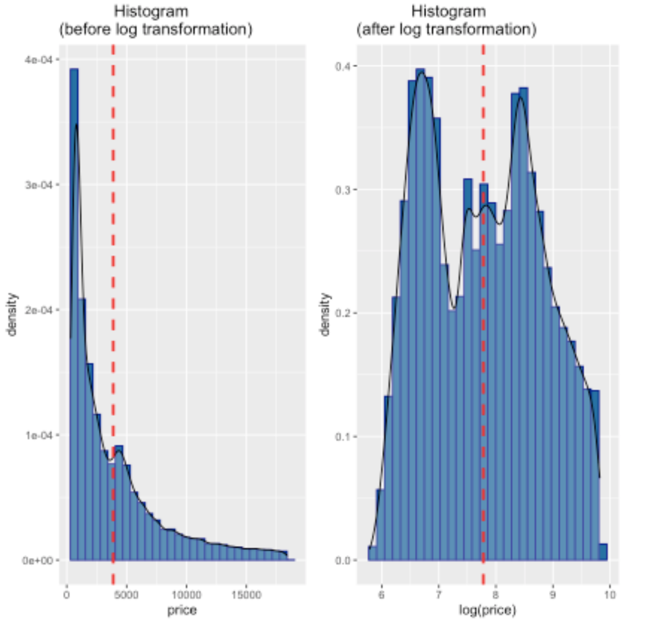
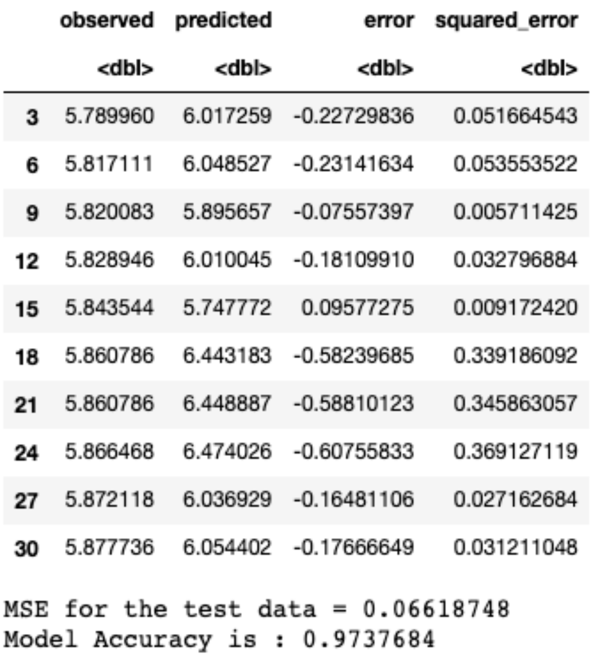


Fig. 9(a). Histogram of 'price'

Fig. 9(b). Histogram of log('price')

4.1) Model Summary – In this multiple linear regression model, the main objective is to check whether the response variable is linearly dependent on the other independent predictor variables. In other words, the *Null Hypothesis* for this model is, *there is no effect of the independent variables on the response variable i.e. all the coefficients of the independent variables are zero.* But from the model summary, it is clear that all the p-values for the coefficients of the corresponding independent variables are significantly smaller (considering the level of significance 0.01) which suggests that there is no statistically significant evidence not to reject the null hypothesis. Hence our model concludes that all the predictor variable has some effect on the response variable 'price'.

4.2) Model Performance and Accuracy – The following figure (Fig. 10) shows the observed values, predicted values, corresponding errors and the squared error values for the first 10 sample observations in the testing set based on the regression model. Now the *Mean Square Error (MSE)* for this model is 0.066 which is significantly low and the model accuracy, calculated by using MAPE (Mean Absolute Percentage Error) measure, is 97%. So our model has high performance in predicting the price of a diamond based on certain physical properties.



	observed	predicted	error	squared_error
	<dbl>	<dbl>	<dbl>	<dbl>
3	5.789960	6.017259	-0.22729836	0.051664543
6	5.817111	6.048527	-0.23141634	0.053553522
9	5.820083	5.895657	-0.07557397	0.005711425
12	5.828946	6.010045	-0.18109910	0.032796884
15	5.843544	5.747772	0.09577275	0.009172420
18	5.860786	6.443183	-0.58239685	0.339186092
21	5.860786	6.448887	-0.58810123	0.345863057
24	5.866468	6.474026	-0.60755833	0.369127119
27	5.872118	6.036929	-0.16481106	0.027162684
30	5.877736	6.054402	-0.17666649	0.031211048

```
MSE for the test data = 0.06618748
Model Accuracy is : 0.9737684
```

Fig. 10. MSE Table

5) Some Observations and Conclusion – Such a high model accuracy often suggests overfitting due to the presence of high variance or sometimes, in the linear regression model the violation of *model assumptions.* Hence I examined all the tests on *model assumptions,* such as the *NCV test (test for homoscedasticity), VIF test (test for autocorrelation)* and *Durbin-Watson test (test of independence of random errors)* and found that this model has failed to justify all the prior assumptions. So in spite of getting a high model accuracy and very low MSE, we can not conclude that this model is good for predicting the price of a diamond based on its physical properties.

IV. Concluding Remarks

The prime objective of this project is to build predictive models that can provide insights in regard to the model performance and application limitations of machine learning methods in different contexts. In this project report, I discuss the outcome of four classification models and a regression model on three different datasets. I apply the KDD methodology to evaluate the model accuracy and the prediction performance on a new sample set. A comparative study is conducted between the models over both the datasets based on specific measures and thus predicts the best-performing model for individual dataset.

In future, it would be interesting to see how an ensemble of these classifier or regressor works on the same dataset and analyze the impact of each component.

References

[1] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics,* vol. 5, no. 4, pp. 115-133, 1943.

[2] PwC, "Protecting the perimeter: THe rise of external fraud," *PwC's Global Economic Crime and Fraud Survey 2022,* pp. 1-15, 2022.

[3] J. A. M. e. al., "An introduction to decision tree modeling," Journal of Chemometrics: A Journal of the Chemometrics Society, vol. 18, pp. 275-285, 2004.

[4] J. R. Quinlan, "Learning decision tree classifiers," ACM Computing Surveys (CSUR), vol. 28, pp. 71-72, 1996.

[5] Y.-Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," Shanghai archives of psychiatry, vol. 27, p. 130, 2015.

[6] T. Chen and C. Guestrin, Xgboost: A scalable tree boosting system, 2016.

[7] A. Ogunleye and Q.-G. Wang, "XGBoost model for chronic kidney disease diagnosis," IEEE/ACM transactions on computational biology and bioinformatics, vol. 17, pp. 2131-2140, 2019.

[8] Y. Qi, Random forest for bioinformatics, Springer, 2012.

[9] A.-L. Boulesteix, S. Janitza, J. Kruppa and I. R. Knig, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 2, pp. 493-507, 2012.

[10] I. Rish, "An empirical study of the naive Bayes classifier," IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, pp. 41-46, 2001.

[11] F.-J. Yang, "An implementation of naive bayes classifier," 2018 International conference on computational science and computational intelligence (CSCI), pp. 301-306, 2018.

[12] N. R. Draper and H. Smith, Applied regression analysis, John Wiley & Sons, 1998.

[13] S. Chatterjee and A. S. Hadi, Regression analysis by example, John Wiley & Sons, 2006.

[14] Brachman, Ronald J. and Tej Anand. “The Process of Knowledge Discovery in Databases: A First Sketch.” *AAAI-94 Workshop on Knowledge Discovery in Databases* (1994): n. pag.