Application Of Machine Learning Techniques In Retail And Airline Industry

Pramod Ramu
StudentId: 20205759
Data Mining and Machine Learning 1, MSc in Data Analytics
National College of Ireland Dublin, IRELAND
Email: x20205759@student.ncirl.ie. URL: www.ncirl.ie

***Abstract*—In recent days machine learning has been incorporated in many industries. This project aims to implement machine learning models on the large data sets of the airline and retail industry. The data sets which were used in this project are the Rossman store, airline customers satisfaction, and the flight price. The aim was to predict the sales, customer satisfaction, and airfare price. This project encompasses multiple regression, logistic regression,k-nearest neighbors, decision tree regression, and random forest regression, models. The project uses Knowledge Discovery in Databases(KDD) methodology. Each model has been evaluated with respective measuring metrics and the results have been predicted.**

***Index Terms*—Machine Learning , KDD , Multiple Regression, Logistic Regression , K-Nearest neighbours, Decision Tree , Random Forest**

I. Introduction

In recent days machine learning is the top trending technology in the world. It has a variety of use cases in different industries. Industries that handle huge numbers of data incorporate machine learning technology. Today finance, retail, medical, transportation, government, and social media sectors are using machine learning. [1]

This project aims to apply machine learning models to large data sets related to the retail and airline industry and verify the accuracy and performance of each model. The project encompasses three large data sets. Among them, two data sets belong to the airline industry and another belongs to the retail industry. Five supervised machine learning models are incorporated to predict the air travelers' contentment, flight price, and the sales of a retail store. All the data sets are obtained from Kaggle. The main intention of this project is to study the application of machine learning techniques in these areas.

Sales and profit are the important key areas of the business. The goal of every business organization, retail store as well as other shop is to improve sales and profit. This is the data set of drug stores called Rossman Stores [2].The goal is to identify the factors that influence the sales and also to predict the same by using them.

Today customer satisfaction is the main goal of all the industries for increasing sales and profits.In order to achieve this standard the organizations should deliver their products or service to create a good customer experience. This data set [3] encompasses customer reviews based on their journey experience with the airlines on various factors. The goal is to predict whether the customer is satisfied or not based on the service offered by the airlines.

Most people travel to various places through airways for business or personal causes. Usually, for the long-distance travel, it would be planned before a couple of months. In this situation, the flight price is an important factor which is needed to be considered. This data set [4] encompasses travel schedule data of different airlines and the objective is to estimate the flight fare price using the factors available in this data.

To start this project a literature survey was made and identified the methodologies used on different data sets which will be covered in the related work section. The next section gives information about the methodology used in the project as well as steps to preprocess the data sets. After this section, the next section is the data modeling and model evaluation section which gives detailed information on evaluating the applied models based on their fits and also encompasses diagnostic tests and measuring metrics. Finally, in the conclusion section, a summary of the entire project has been provided and in future work, the section gives information on what could be done to improve this project in a better way. The reference section includes all the references that were used in this report.

II. Related Work

Before starting the project a literature survey was made to understand the application of machine learning algorithms on different data sets.As this project focus on data set related retail and airline industry so focus was on to identify the papers on related to this area to understand the methodologies used and to summarize the positive and negative aspects of it.

Akshay Krishna, Akhilesh V, Animikh Aich, Chetana Hegde [5] used multiple regression, polynomial Regression, ridge regression, lasso regression, AdaBoost, xgboost for predicting sales. They found that have found that the applying boosting models have better accuracy with an R2 value of 0.59 than the regression models but it would be better if prediction results were included.

A paper published by Ranjitha P and Spandana M used [6] xgboost, linear regression, polynomial regression, and ridge

regression techniques for forecasting the sales of a business such as Big-Mart. They found that model performed better than existing models. There was not much explanation on the data preprocessing.

Another paper published by Yi Zuo, Katsutoshi Yada, A.B.M. Shawkat Ali [7] used linear discriminant analysis, logistic regression analysis,baye classifier, and support vector machine and got the highest accuracy for the support vector machine which was around 90.63 percent.

In another paper, Sunil K Punjabi, Vikyhat Shetty, Shreemun Pranav, Abhishek Yadav [8] used polynomial regression for predicting sales. The accuracy was 0.9016 on applying the polynomial regression. Another model could be applied to get better clarity on the accuracy and the it could be compared to evaluate the performance of the models.

Tianyi Wang, Samira Pouyanfar, Haiman Tian, Yudong Tao, Miguel Alonso, Steven Luis Shu-Ching Chen [9] used linear regression, support vector machine,xgboost, and random forest for predicting flight delay in aviation dataset. The highest accuracy was found on applying the random forest model than other models which was around 86 percent.

In another paper published by Praphula Kumar Jain; Rajendra Pamula; Sarfraj Ansari; Dilip Sharma; Lakshmibai Maddala [10] predicted the airline based on the customer-generated feedback data. The models used were k nearest neighbor, support vector machine, and decision tree. Support vector machine model had the highest accuracy of 82.75 percent.

In another paper, Navoneel Chakrabarty [11] used a gradient boosting classifier model was used to predict the flight arrival delay using a gradient boosting classifier model. The accuracy of the model was 85.73 percent.

Airline passenger load was predicted by Ma Nang Laik, Murphy Choy, Prabir Sen [12] using the decision tree model. The accuracy was around 96 percent but the model evaluation steps would have been much more detailed.

In another paper Rahul Nigam, K. Govinda [13] used a logistic regression model to predict whether the flight will be delayed or not. Even though the accuracy was 80.6 percent there was not much detailing regarding the data preprocessing and evaluation of the model.

In another paper K. Tziridis, Th. Kalampokas, G. A. Papakostas, K. I. Diamantaras [14] used regression tree, random forest, bagging regression tree, support vector machine both linear and polynomial to predict airfare prices. The highest accuracy was found in the random forest regression tree model which was around 79.49 percent and the lowest was support vector machine linear which was around 44.95 percent.

Yingchao Xiao Yuanyuan Ma, Hui Ding [15] used k-nearest neighbors to predict the air traffic flow. The objective was to showcase that the model performs better than the support vector machine but the evaluation steps was not much clear.

To perform opinion mining on US Airline twitter data Abdelrahman I. Saad used support vector machine, logistic regression, random forest,xgboost,naïve Bayes, and decision tree models. Support Vector Machine had the highest accuracy which was 83.31 percent and decision tree had the lowest accuracy which was around 70.51 percent.

```
$ Store         : int
$ DayOfWeek     : int
$ Date          : chr
$ Sales         : int
$ Customers     : int
$ Open          : int
$ Promo         : int
$ StateHoliday  : chr
$ SchoolHoliday: int
```

Fig. 1. Rossman store Dataset

III. Methodology

The field of data science mainly involves the analysis of a huge amount of data to retrieve useful outcomes from it. To carry out this process, certain methodologies have been incorporated. The most popular methodologies are Knowledge Discovery in Databases(KDD), Cross-Industry Standard Process(CRISP-DM), and Sample Explore Modify Model Assess(SEMMA).CRISP-DM encompasses business understanding, data understanding, data preparation, modeling, evaluation, and deployment. As the name suggests SEMMA includes sample, explore, modify, model, and assess stages.KDD is a bit different than the other two methodologies. It mainly involves data cleaning, data integration, data selection, data transformation, data mining, evaluation, and visual representation of the obtained knowledge. This project follows the KDD methodology as the desired outcome to represent the knowledge obtained. So in all the further sections the process is showcased based on the KDD methodology. [16]

A. Data set Selection

As mentioned earlier all the data sets were retrieved from Kaggle. This project encompasses three large data sets. The first data set is the 'Rossman store' data set figure1, which encompasses data of a drug store. Here the dependent variable is sales which is a continuous variable. When the data set was loaded it had 1017209 observations and later it was reduced to one lakh for working purposes. It had nine columns and later on doing transformations the number of columns reached eleven. The second data set is airline customer satisfaction figure2 which consists of 129880 observations with 23 columns.Here the response variable is satisfaction which is a dichotomous variable that is whether the customer is satisfied or dissatisfied. TThe third data set is the flight price data set

```
$ satisfaction                          : chr
$ Gender                                : chr
$ Customer.Type                         : chr
$ Age                                   : int
$ Type.of.Travel                        : chr
$ Class                                 : chr
$ Flight.Distance                       : int
$ Seat.comfort                          : int
$ Departure.Arrival.time.convenient: int
$ Food.and.drink                        : int
$ Gate.location                         : int
$ Inflight.wifi.service                 : int
$ Inflight.entertainment                : int
$ Online.support                        : int
$ Ease.of.Online.booking                : int
$ On.board.service                      : int
$ Leg.room.service                      : int
$ Baggage.handling                      : int
$ Checkin.service                       : int
$ Cleanliness                           : int
$ Online.boarding                       : int
$ Departure.Delay.in.Minutes            : int
$ Arrival.Delay.in.Minutes              : int
```

Fig. 2. Airline Customer Satisfaction

```
$ Airline          : chr
$ Date_of_Journey: chr
$ Source           : chr
$ Destination      : chr
$ Route            : chr
<U+2192> BOM <U+2192>
$ Dep_Time         : chr
$ Arrival_Time     : chr
$ Duration         : chr
$ Total_Stops      : chr
$ Additional_Info: chr
$ Price            : num
```

Fig. 3. Flight Price Prediction

figure3. It consists of 10683 observations with 11 columns. Here the dependent variable is the price which is a continuous variable. The goal is to predict the flight price.

IV. DATA PREPROCESSING

As per the KDD methodology, the first step is data preprocessing. To build a model, it is necessary to preprocess the data. Sometimes the data set will be unstable, the preprocessing helps to make it stable. It mainly involves data cleaning, data transformation, data wrangling, encoding the categorical variables, exploratory data analysis, and data splitting.

A. Rossman Store : Dataset Preprocessing

The process started by loading the data set. As mentioned earlier it had 1017209 observations with 9 columns. The data set was tested to check the presence of missing values(NA values). It was observed that there were no missing values. After checking the structure of the data set it was observed that the 'Date' and 'StateHoliday' column was in character type. The date column was split into three columns such as 'Year' 'Month' and 'Day' by extracting their respective numerical values. Initially, the values in these three columns were in character and it was converted to numeric type and the 'Date' variable was dropped. The 'StateHoliday' had categorical variables '0', 'a',' be, and 'c', and it was encoded using label encoding. All the columns were converted numeric type. Since it was a huge data set it was reduced to 117210 observations for working purposes. It was observed that the 'year' variable consists of same value through out,so it was dropped from the data set.Once the data set was reduced the data set was rechecked and to verify whether it is stable or not. Once it was confirmed that the new filtered data set is stable then it was split into training and test sets with 80 percent observations given for the training set and 20 percent for the test set. The training set consists of 94664 observations and the test set consists of 22546 observations.

B. Rossman Stores : Data Modelling

Since the dependent variable was a continuous variable, it was decided to use multiple regression model.This process incorporated backward elimination method means removing the variables one by one based on highest p-value(not statistically significant).Initially all the independent variables were included along with the dependent variable and the target data was the training set.It was observed that 'Month' and 'SchoolHoliday' variables had highest p-values among that the p-value of "SchoolHoliday" was more and it was removed.The model was ran again and this time 'month' variable had the highest 'p' value and it was removed from the model.The model was ran again and it was observed that all the variables were statistically significant and the adjusted R-square was 0.8478

Rossman Stores : Model Evaluation

The first thing is to make sure whether the model satisfies all the assumptions of the multiple regression such as linearity, homoscedasticity, no auto-correlation, absence of multicollinearity, no influential data points, and normal distribution. When the model was plotted it was observed that the homoscedasticity assumption was not up to the mark. To fix this issue the thumb rule is to take the log or square root transformation of the dependent variable. On taking the log, it was not observed much changes, so square root transformation was applied on the dependent variable of the training set and

```
Call:
lm(formula = sqrt(Sales) ~ Store + DayOfWeek + Customers + Open +
    Promo + StateHoliday + Day, data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max
-85.858  -4.790  -0.075   4.541  92.985

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.887e+00  1.723e-01   28.36   <2e-16 ***
Store        -1.088e-03  9.047e-05  -12.02   <2e-16 ***
DayOfWeek    -7.024e-01  1.883e-02  -37.31   <2e-16 ***
Customers     3.138e-02  7.857e-05  399.36   <2e-16 ***
Open          5.283e+01  1.177e-01  448.81   <2e-16 ***
Promo         8.908e+00  6.660e-02  133.76   <2e-16 ***
StateHoliday -6.508e+00  1.839e-01  -35.38   <2e-16 ***
Day           3.746e-02  3.335e-03   11.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.95 on 94656 degrees of freedom
Multiple R-squared:  0.9356,    Adjusted R-squared:  0.9356
F-statistic: 1.964e+05 on 7 and 94656 DF,  p-value: < 2.2e-16
```

Fig. 4. Model for predicting sales

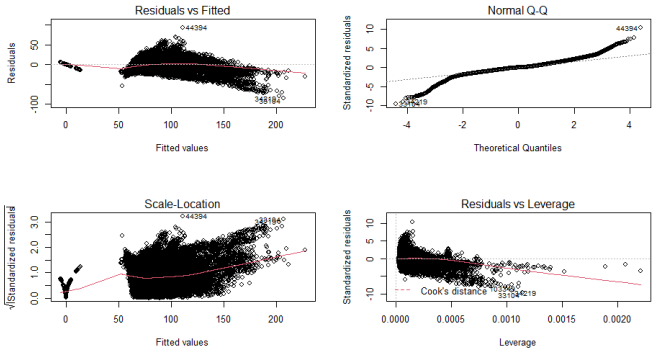

Fig. 5. Plot of the applied model to predict sales

```
> vif(first_model3)
    Store DayOfWeek Customers      Open     Promo StateHoliday
 1.000669  1.674452  1.642380  2.323497  1.240252     1.353474
>
```

Fig. 6. VIF Test

```
> durbinWatsonTest(first_model3)
 lag Autocorrelation D-W Statistic p-value
   1      0.04988347       1.900222       0
 Alternative hypothesis: rho != 0
```

Fig. 7. Durbin-Watson Test

```
> influencePlot(model = first_model3,scale = 3,main = "Influence Plot")
        StudRes          Hat        CookD
33104  -9.603124 0.0011400796 0.0131446003
34219  -8.961141 0.0010449940 0.0104915779
44394  10.396077 0.0001486875 0.0020067676
87232  -1.754204 0.0020726151 0.0007988783
101727 -3.518172 0.0022141411 0.0034328892
```

Fig. 8. Cook's D values

	actuals	predicteds
2	72.54654	83.90711
4	91.18114	111.33986
5	90.74139	81.82976
6	85.28189	82.77146
7	95.34674	108.69684
8	109.29776	90.43725

Fig. 9. Actual vs Predicted values

it was applied to the test set as well to obtain the proper predictions. On applying the square root transformation to the dependent variable, the model was run again. The results were satisfactory the adjusted R-square value at this stage was 0.9355

From the model figure4 it was observed that the R-square and adjusted r square values which are measured with the scale of 0 to 1 were equal which was 0.9355. So 93.55 percent of the variation in the dependant variable could be explained by the model. The overall p-value was statistically significant. The residual standard error was observed at 8.95 for 94656 degrees of freedom. Simultaneously few diagnostic tests were conducted to verify whether the assumptions are satisfied or not. From the plot figure5 it was observed that the model satisfies linearity and normal distribution which was observed from residuals vs fitted plot and normal Q-Q plot but there was some mild variance as observed from the scale-location plot. To verify the absence of multicollinearity variance inflation factor(vif) test was conducted figure6. The values should not be more than 5 for a good model and the obtained results were satisfactory. Durbin-Watson test figure7 was conducted to verify the autocorrelation between the explanatory variables. The Durbin-Watson statistic should be within 2 for a good model. The observed value was 1.90. The influential data points were checked by verifying Cook's distance figure8 and also from the model plot figure5 and the results were satisfactory.

On completing the diagnostic tests the next step was to predict the sales by applying the model on the test set. On predicting the sales, it was observed that there was some mild variation in the actual and predicted values figure9.

Airline Customer Satisfaction : Data Preprocessing

The data cleaning was the first step to start as per the thumb rule of the KDD process. When the data were checked for missing data the 'Arrival.Delay.in.Minutes' had 393 missing values. In order to handle the missing values the relationship between the 'Arrival.Delay.in.Minutes' and 'Depar-

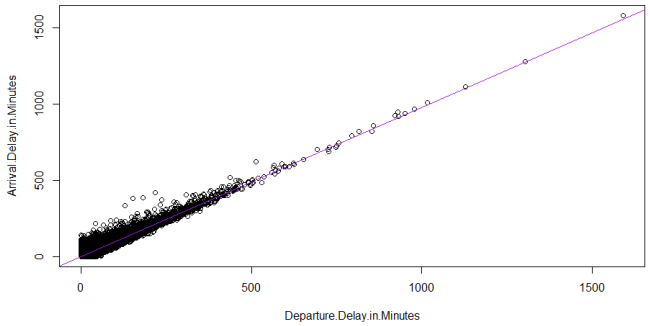

Fig. 10. Arrival Delay in Minutes vs Departure Delay in Minutes

```
Model: "glm, satisfaction ~ ., binomial, training_set2"
Null:  "glm, satisfaction ~ 1, binomial, training_set2"

$Pseudo.R.squared.for.model.vs.null
                              Pseudo.R.squared
McFadden                              0.434190
Cox and Snell (ML)                    0.450100
Nagelkerke (Cragg and Uhler)          0.601943

$Likelihood.ratio.test
 Df.diff LogLik.diff Chisq p.value
     -22      -27185 54370       0
```

Fig. 11. Nagelkerke Method

ture.Delay.in.Minutes' was checked. It was observed that there was a linear relationship between these two variables figure11.

So the missing values were removed by imputing the mean of 'Arrival.Delay.in.Minutes'.So the next step was transforming the variables to the required format and type. It was observed from the data set that 'satisfaction', 'Gender', 'Customer. Type', 'Type.of.Travel', 'Class' were categorical variables with 2 and 3 values.It was encoded with label encoding. The 'satisfaction is the dependent variable and the outcome seems to be classified, so it was transformed to factor type with the other categorical variables 'satisfaction', 'Gender', 'Customer. Type' and 'Type.of.Travel'. The rest of the variables were kept as integer types. Once this process was completed the data set was rechecked and also data wrangling operations were implemented.Now the data set was split into training and test sets where 70 percent of the observation was given to the training set and 30 percent to the test set.The training set had 90916 observations and the test set had 38964 observations.It was observed that the values of 'Age','Flight.Distance','Departure.Delay.in.Minutes', 'Arrival.Delay.in.Minutes' were bigger when compared to other values.So it was scaled using the feature scaling method for both training and test sets. The training and test sets were rechecked before starting to build the model.

Airline Customer Satisfaction : Data Modelling As the dependent variable is dichotomous, it involved classification models. So it was decided to use logistic regression and k-nearest neighbors machine learning models. A first logistic regression model was applied to the training set. Initially, the model was ran including all the independent variables.Once the model was ran it was observed that the 'Class' variable was not statistically significant. So it was dropped and the model was ran again including all the other independent variables except 'Class' variable. It was observed that all the independent variables were statistically significant. The K-nearest neighbors model was applied using the ideal value of K.The k-value is usually the square root of the number of observations. The observations found in the training set are 90916. So the K-value was approximately 302, as per the thumb rule.

Airline Customer Satisfaction : Model Evaluation As all the variables in the logistic regression model were statistically significant, the next step was to measure the level of goodness of the fit. This is called the pseudo-R-squared value. This was verified using the Nagelkerke method. Here the pseudo-R-squared value should not be statistically significant that is the value should be greater than 0.05. It was observed that the pseudo R square value was 0.602 figure11. After verifying the pseudo R square the next step was the prediction. The model was applied to the test set for predicting customer satisfaction. Later the prediction was tested using the confusion matrix.The accuracy,kappa,sensitivity and specificity values were 0.8351,0.6674,0.8197 and 0.8479 figure12.The ROC curve has been showing how well the model has been fit. By observing the roc curve the model seems to be good and the AUC value was 0.9075.

After observing the results from the logistic regression model, now K-nearest neighbors model was applied by taking the k value 301 which is one less than the ideal value. Cross table has been implemented to verify the results. From the confusion matrix, the accuracy observed was 0.9219, and the kappa, sensitivity, and specificity values were 0.8431,0.9426, and 0.9047 respectively. The model was also ran using other nearest k-values.The ROC the curve has been implemented for the moel.The AUC value was 0.9236. Based on accuracy among the logistic regression and knn model,the knn model fits best for the data set.

Flight Price : Data Preprosessing

Here the data set was already split into two sets. The data set with more observations were chosen. The size of this data set is small when compared to the previous two data sets. It had 10683 observations with 11 columns. The process started by cleaning the data set. It was observed that the 'Route' and 'Total-Stops' had missing values. So to the rows with 'Route' and 'Total-Stops' with missing values were removed. The data set was rechecked and there were no missing values. So the next was to transform the variables to the required format. The 'DateofJourney' was a character variable in 'dd/mm/yy format. So it was split into 'traveling day' and 'traveling month' by extracting the numerical values of their respective categories. Initially, it was character type and it was converted to numerical type, the year variable was dropped since it

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 14457  3243
         1  3181 18083

               Accuracy : 0.8351
                 95% CI : (0.8314, 0.8388)
    No Information Rate : 0.5473
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6674

 Mcnemar's Test P-Value : 0.4466

            Sensitivity : 0.8197
            Specificity : 0.8479
         Pos Pred Value : 0.8168
         Neg Pred Value : 0.8504
             Prevalence : 0.4527
         Detection Rate : 0.3710
   Detection Prevalence : 0.4543
      Balanced Accuracy : 0.8338

       'Positive' Class : 0
```

Fig. 12. Confusion Matrix Aand Statistics For Logistic Regression Model

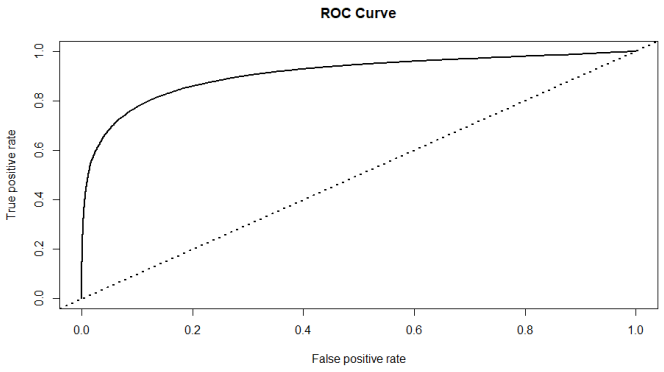

Fig. 13. ROC Curve for Logistic Regression Model

```
> rc_pred_lr <- predict(second_Model, newdata=test_set2, type="response")
> rc_pr <- prediction(rc_pred_lr, test_set2$satisfaction)
> rc_prf_lr <- performance(rc_pr, measure = "tpr", x.measure = "fpr",auc=TRUE)
> plot(rc_prf_lr,main = "ROC Curve",col = 1,lwd = 2)
> abline(a = 0,b = 1,lwd = 2,lty = 3,col = "black")
> auc1 <- performance(rc_pr, measure = "auc")
> auc1 <- auc1@y.values[[1]]
> auc1
[1] 0.9075516
```

Fig. 14. AUC for logistic Regression Model

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  38964

                           | third_model
test_set2$satisfaction     |         0 |         1 | Row Total |
---------------------------|-----------|-----------|-----------|
                         0 |     16625 |      1013 |     17638 |
                           |     0.943 |     0.057 |     0.453 |
                           |     0.891 |     0.050 |           |
                           |     0.427 |     0.026 |           |
---------------------------|-----------|-----------|-----------|
                         1 |      2030 |     19296 |     21326 |
                           |     0.095 |     0.905 |     0.547 |
                           |     0.109 |     0.950 |           |
                           |     0.052 |     0.495 |           |
---------------------------|-----------|-----------|-----------|
                  Column Total |     18655 |     20309 |     38964 |
                           |     0.479 |     0.521 |           |
---------------------------|-----------|-----------|-----------|
```

Fig. 15. Cross Table For KNN Model

```
Confusion Matrix and Statistics


third_model     0     1
          0 16625  2030
          1  1013 19296

               Accuracy : 0.9219
                 95% CI : (0.9192, 0.9245)
    No Information Rate : 0.5473
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8432

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9426
            Specificity : 0.9048
         Pos Pred Value : 0.8912
         Neg Pred Value : 0.9501
             Prevalence : 0.4527
         Detection Rate : 0.4267
   Detection Prevalence : 0.4788
      Balanced Accuracy : 0.9237

       'Positive' Class : 0
```

Fig. 16. Confusion Matrix And Statistics For KNN Model

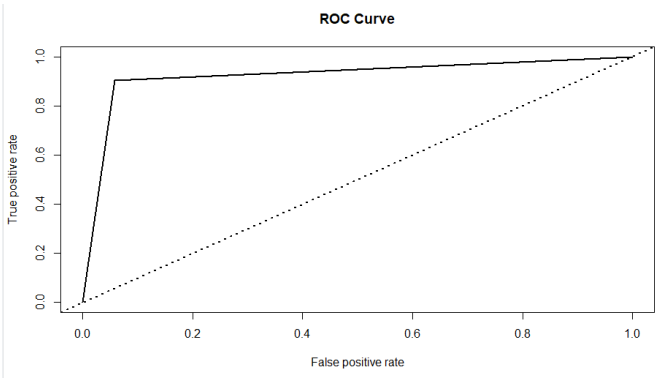

Fig. 17. ROC Curve for KNN Model

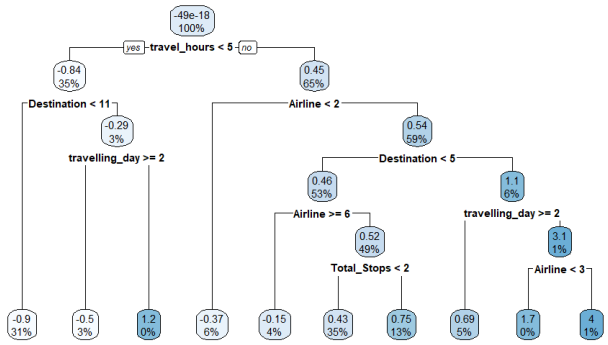

Fig. 18. Decision Tree For Flight Price Prediction

was the same year for all and the 'DateofJouney' variable was dropped. The 'DepTime' 'ArrivalTime' variables were the combination of hours and minutes and it was in character type. These variables were split into 'departure hours' 'departure minute' 'arrival hours' and 'arrival minutes' by extracting their respective numerical values and the type was also converted from character to numerical and the 'DepTime' and 'Arrival-Time' variables were dropped. The duration variable was also a string variable with the combination of hours and minutes. So the numerical values were extracted to a separate column called 'travel hour' and 'travel minutes' and the type was also converted from character to numerical and the duration variable was dropped. It was observed that the 'Airline', 'Source', 'Destination', 'Total Stops' were categorical variables.' Airline' variable encompasses a list of airline names, 'Source' variable consists of the names of the cities where the flight take off, the 'Destination' variable consists of the names of the cities where the flight land, and the 'Total Stops' consists of the values of several stops. So these variables were encoded using label encoding and were converted to numerical type. The 'Route' variable was dropped since it had the same information in 'Source', 'Destination'.The 'Additional Info' variable was dropped as it didn't have much information. Once the transformation process was completed, the data set was rechecked and the data set was split into training and test sets.70 percent of the observations were given to the training sets and 30 percent of the observations were given to the test set. The training set consists of 7617 observations and the test set encompasses 3065 observations. As the values of the 'Price' variable were larger it was scaled using feature scaling methods.

Flight Price : Data Modelling

Once the data preprocessing was completed. It was decided to apply regression tree models for the prediction as the dependent variable was the continuous variable. So decision tree and random forest models were applied to the training set. The first decision tree model was applied by taking 'Price' as the dependant variable and all the other variables as the independent variables and later the same process was followed by the random forest model.

	actuals	predicteds
1	1.136847106	0.7500474
2	-0.001593015	0.4258238
3	0.308474596	0.4258238
4	1.022202181	0.7500474
5	-0.448273441	0.6913704
6	-0.264750028	0.4258238

Fig. 19. Actual and Predict Values Of The Decision Tree

Flight Price : Model Evaluation The tree plot for the applied decision tree model is shown in the figure17. The next step was to predict the 'Price' values using this model. The mean square error was around 0.348 and the root mean square error was around 0.59.The accuracy was around 66 percent. The actual and predicted values has been shown in the figure18

So basically the decision tree is built using all the explanatory variables of the training data set or specific independent variables whereas in the random forest it randomly selects the observations and the variables for building multiple decision trees and then the average is taken to get the final result.So on running the random forest. The random forest model was applied on keeping the 'Price' variable as the dependent variable and the other variables as the predictors. The plot of the model is shown in the figure19. Once the model was built then it was used to predict the 'Price' values of the test set. The actual and predicted values on applying random forest model is shown in the figure20

The mean square error was around 0.15 and the root mean square error was around 0.39.The accuracy of the model was around 85 percent. So based on the accuracy random forest model fits best for the data set.

V. Conclusions and Future Work

Overall 5 supervised machine learning algorithms were applied on three large data sets. In the first data

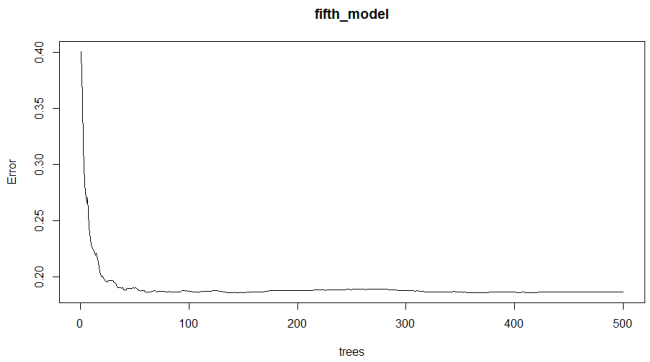

Fig. 20. Plot For Random Forest Model

	actuals	predicteds
1	1.136847106	0.88938755
2	-0.001593015	0.02912865
3	0.308474596	0.24693733
4	1.022202181	0.87569103
5	-0.448273441	-0.34547817
6	-0.264750028	0.49631323

Fig. 21. Actual And Predicted Values Using Random Forest Model

set(Rossman store) multiple regression algorithm was applied and almost all the assumptions were satisfied and able to see a mild variance. Also, satisfactory results were obtained. So in the future, these results could be improved by applying boosting algorithms and would be interesting to study the results. In the second data set (Airline customer satisfaction) it was observed that the k-nearest neighbor model had better accuracy than the logistic regression model and in the future results could also be studied by applying random forest and decision tree classification algorithms. In the third data set (flight price) as the dependent variable was a continuous variable regression tree models were incorporated. It was observed that the results of regression decision trees were much quicker than the random forest model but in terms of accuracy, the random forest model was more accurate.

REFERENCES

[1] "Uses of machine learning," https://www.sas.com/en_ie/insights/analytics/machine-learning.html, accessed: 2021-12-13.

[2] "Rossman store," https://www.kaggle.com/c/rossmann-store-sales, accessed: 2021-11-4.

[3] "Airline customer satisfaction," https://www.kaggle.com/sjleshrac/airlines-customer-satisfaction, accessed: 2021-11-7.

[4] "Flight price," https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh, accessed: 2021-11-7.

[5] A. Krishna, A. V, A. Aich, and C. Hegde, "Sales-forecasting of retail stores using machine learning techniques," in *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, 2018, pp. 160–166.

[6] R. P and S. M, "Predictive analysis for big mart sales using machine learning algorithms," in *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2021, pp. 1416–1421.

[7] Y. Zuo, K. Yada, and A. S. Ali, "Prediction of consumer purchasing in a grocery store using machine learning techniques," in *2016 3rd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, 2016, pp. 18–25.

[8] S. K. Punjabi, V. Shetty, S. Pranav, and A. Yadav, "Sales prediction using online sentiment with regression model," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2020, pp. 209–212.

[9] T. Wang, S. Pouyanfar, H. Tian, Y. Tao, M. Alonso, S. Luis, and S.-C. Chen, "A framework for airfare price prediction: A machine learning approach," in *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, 2019, pp. 200–207.

[10] P. K. Jain, R. Pamula, S. Ansari, D. Sharma, and L. Maddala, "Airline recommendation prediction using customer generated feedback data," in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, 2019, pp. 376–379.

[11] N. Chakrabarty, "A data mining approach to flight arrival delay prediction for american airlines," in *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*, 2019, pp. 102–107.

[12] M. N. Laik, M. Choy, and P. Sen, "Predicting airline passenger load: A case study," in *2014 IEEE 16th Conference on Business Informatics*, vol. 1, 2014, pp. 33–38.

[13] R. Nigam and K. Govinda, "Cloud based flight delay prediction using logistic regression," in *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, 2017, pp. 662–667.

[14] K. Tziridis, T. Kalampokas, G. A. Papakostas, and K. I. Diamantaras, "Airfare prices prediction using machine learning techniques," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1036–1039.

[15] Y. Xiao, Y. Ma, and H. Ding, "Air traffic flow prediction based on k nearest neighbor regression¡sup¿*¡/sup¿," in *2018 13th World Congress on Intelligent Control and Automation (WCICA)*, 2018, pp. 1265–1269.

[16] U. Shafique and H. Qaiser, "A comparative study of data mining process models (kdd, crisp-dm and semma)," *International Journal of Innovation and Scientific Research*, vol. 12, no. 1, pp. 217–222, 2014.