Prediction of Online News Popularity, Reused Cars Price, Bank Marketing Terms Client Acceptance Using Data Mining Algorithms

Kamesh Munuswamy
x19199562
MSc Data Analytics
National College of Ireland
Dublin, Ireland

***Abstract*—This paper investigates the widely used data mining models to predict the online news popularity (Dataset-1), forecasting the resale prices of used cars (Dataset-2) and classifying the significant factors of bank customers relationship in term acceptance (Dataset-3). First, comparative study of various prediction methods provides the evidence to use Linear regression and classification models like Logistic regression, Random forest, K-Nearest Neighbors (KNN), Decision tree. Data from three different sectors are analysed and being subjected to machine learning algorithms, to classify and predict the behaviour; Ultimately, data are used for decision making. This study uses the data mining techniques to provide improved linear and classified predictions, which demonstrates optimal observation threshold to forecast accurate result as possible. The outcome of this paper is to provide the best possible forecast of online news popularity, to estimate accurate resale price of used cars and to understand the customers relationship in banking market.**

Key words: Linear Regression, Logistic Regression, K-nearest neighbouring (KNN), Decision Tree, Random Forest, Online News Popularity, Reused Cars Price and Bank Customer Term.

I. Introduction

Dataset-1: Online News popularity

In recent times, with the widespread use of smartphones, data transmission (e.g. news, video clips, articles and images) has played an increasingly important role in digital news popularity. Note that online news content includes a number of key properties. For example, it is feasibly produced and small in size; its lifespan is short and its expense is low. These features make news content more efficient for social media platforms. More amusingly, this type of content can attract the attention of a drastic number of Internet users within a short period of time. As a matter of fact, researchers concentrate on examining online news information such as forecasting the relevance of news articles, illustrating the decrease in interest over time to understand the online news world because it has so many wider implications [1].

Tatar et al. [3] have demonstrated two types of popularity prediction approaches that are After-publication: a more common approach that utilizes elements that capture the attention that one article receives after its publication. Before-publication: a relatively challenging and productive strategy. This tactic uses only the web metadata features that are identified prior to the publication of the web instead of using related features that one user gets after the release of the content. Previously, researches trying to predict the popularity of the online news and article by forecasting the number of shares. However, in this study we will classify and predict the online news popularity by categorizing them based on the number of shares. Hence, based on the analysis carried out, we use KNN and Random forest classification models to predict the online news popularity.

Dataset-2: Reused Cars Price Prediction

This report discusses the ability of forecasting approaches to improve decision-making in the automobile industry. More precisely, we focus on the second-hand car market and construct the Linear Regression and Random Forest Model for forecasting resale prices. In this study we use Craigslist of US region data which is of the largest collector of used cars for sale. It includes most of the related information that Craigslist offers on car sales, including columns such as size, condition, seller, latitude / longitude, and 18 additional categories.

According to data gathered from the Agency for Statistics of BiH, 921,456 vehicles were registered in 2014, of which 84% are cars for personal use. This figure is risen by 2.7% since 2013 and this phenomenon is likely to persist and the number of cars will grow significantly in future [4]. In practical, making solid pricing judgments requires a clear estimation of demand. In the second-hand car industry, demand lies exponentially on the disparity between the residual value of the car and its sale price [5]. Apparently, in this study prediction of used cars price is carried out by considering all the important aspects of second-hand car features like year of build, manufacturer, model, condition of the car, engine cylinder type, fuel consumed, odometer readings and many more components.

Dataset-3: Bank Marketing Terms Client Acceptance

Banks store a massive amount of data about their clients in daily life. These data are used to build and maintain a consistent engagement with customers in order to target them

individually for specific product or banking offers. Normally, the chosen consumers are approached directly through: personal contact, mobile phone, email, or some other correspondence to promote or deliver a new product / service, this form of marketing is called direct marketing. In reality, direct marketing is one of the key tactics of several banks and insurance companies to associate with their customers [6]. There are two main paradigms through which organizations promote their products and services, i.e. by widespread crusades that focus on the community as a whole and direct campaign, with targeted initiatives that target only a specific group of individuals [7].

In this paper, for predicting the client marketing term acceptance, we have used the most widely practiced data mining classification models such Logistic regression and Decision tree. The dataset is based on Bank marketing- UCI dataset. The data corresponds to the direct marketing campaigns of the Portuguese banking institution. The marketing strategies were focused on telephone calls. Also, more than one interaction with the same client was expected to be accessed whether the product (bank term deposit) was ('yes') or not ('no') subscribed. The classification objective is to determine whether the client will accept a term deposit (variable y).

The remainder of the paper is organized as follows: Section , narrates the related works on this subject. Section , explains the data mining methodology used in this study. Section , in this part data mining models which is being used is defined. Section , results of the data mining models are evaluated and explained. Section , finally conclusion.

Research Question

RQ1: *"Predicting the popularity of the online news published by using the target variable shares."*

RQ2: *"To predict the reused cars price using multiple independent factors."*

RQ3: *"The aim is to predict whether the client will sign a term (y variable- yes/no)."*

II. Related work

Dataset-1: Online News popularity

There are several researches that propose various schemes to forecast the popularity of different online content.

Md. Taufeeq Uddin et al. (2016) [1] emphasizes on the popularity forecast of digital news by forecasting whether or not people share an article and how many users share pre-publication content. This journal recommends a gradient booster model (GBM) utilizing features that are identified prior to article published. The GBM can measure popularity using only statistical attributes associated with the original news stories without using the actual text of news articles or at the time of publishing. In this research, Random forest (RF) model is also implemented. When comparing the result of RF and GBM, it is showcased that GBM was better in predicting the popularity of the online news since MAPE LOG and MAPE produced from GBM were 8.11% and 69.42% which was comparatively smaller than RF.

Huangqing Chen et al. (2015) [8] In this paper the author has proposed an improved logarithmic linear prediction system, using an optimum observation threshold to increase prediction efficiency. And the proposal is contrasted with conventional rank algorithms using machine learning techniques. In this study, two simulation models are proposed, such as linear log and a constant scaling model. Due to its superior efficiency, we chose the linear log model with the optimum threshold as our prediction scheme. Our prediction approach is stronger and has a major benefit compared to learning to distinguish algorithms by NDCG [2]. Disadvantage of this paper is that linear log model scheme provides a bad performance due to its different titles characters.

Arapakis et al. (2014) [9] noted out that the estimation of news popularity in the cold start is always a problem. They estimated tweet numbers and page hits using time, news, genre, Wikipedia and Twitter-related features. They mentioned in their findings that the imbalanced degree distribution guided the prediction model to forecast unfavourable articles that inferred the predictions that were not useful in practical scenarios.

Lee et al. (2012) [10] proposed the idea that online content would be common for assessing survival using Cox's proportional hazard regression. They use a variety of measurable parameters to predict and estimate quantitative variables such as the life time of threads and the variety of writings.

Kelwin Fernandes et al. (2015) [12] the paper suggests a modern and constructive Intelligent Decision support program (IDSS), which would analyse articles before publishing. The IDSS initially forecasts how an article would become prominent by using a wide variety of features extracted (e.g., keywords, digital media content, early publication in the article). In this journal, many data mining models were used in prediction such as Random Forest (RF), Adaptive Boosting (AdaBoost), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Na¨ıve Bayes (NB). A Random Forest with a biased control of 73% has achieved the best outcome.

Dataset-2: Reused Cars Price Prediction

Stefan Lessmann et al. (2017) [5] Discusses analytical methods for the estimation of reselling value of used cars. An empirical analysis is carried out to investigate the impact of various degrees of freedom to predict the accuracy in the modelling framework. Next, a comprehensive study of alternate prediction approaches provides the proof that random forest regression is especially good in the forecasting of resale prices. One drawback of complex forecasting is that they increase the dimensions. Although, an analysis of high and low-dimensional projections showed that many methods of forecasting are well suited for large numbers of correlation coefficients. The research result reveals that variability exists in RPF but it will not lower the predictive precision of most advanced predictive methods. Altogether, ensemble methods were observed to provide the most reliable predictions in

different test conditions.

Pudaruth (2014) [13] developed numerous machine learning algorithms for the forecasting price of cars in Mauritius such as: k-nearest neighbors, multiple linear regression analysis, Decision trees and naïve bayes. For less than 1 month the data sets for the development of a forecast model have been gathered manually from local newspapers considering the time factor having a huge effect on the car prices. He explored the following features in his model: brand, model, cubic capacity, miles per kilometre, year of manufacture, exterior colour, transmission type and cost. Nevertheless, the researcher observed that Naive Bayes and Decision Tree was unable to forecast and distinguish quantitative values. In comparison, a small number of dataset instances did not provide strong classification efficiency, i.e. accuracy of less than 70%.

Wu et al. (2009) [14] carried out the forecasting analysis of car price using a skill-based neuro-fuzzy method. The expert program ODAV (Optimal Distribution of Auction Vehicles) has been developed to offer an insight on best price for cars and the place where the best prices will be obtained. To estimate the car price, a Regression model based on the K-nearest neighbors (KNN) machine learning algorithm was used. This method was highly successful, as it has traded more than 2 million cars.

Noor and Jan (2017) [15] constructed a model for the estimation of car prices by means of multiple linear regressions. The two-months dataset included: costs, volume of cubic, paint, date of manufacture, number of ads, power steering, miles per kilometre, method of transmission, engine speed, area, town council, size, version, make. The authors only contemplated engine size, quality, model year and model as input functions. The author manages to produce the accuracy of 98%.

Dataset-3: Bank Marketing Terms Client Acceptance

Hany A. Elsalamony (2014) [6] incorporates the study and implementation of the most significant data mining strategies; multiple perception neural network (MLPNN), tree augmented Naïve Bayes (TAN) and Ross Quinlan new decision tree model (C5.0). The C5.0 model with the accuracy of the maximum values for training data is 94.92% and 93.23% for test samples. Compared to research findings shown by effective models. C5.0 performed considerably better than MLPNN, LR and TAN.

Noor, K el al. (2017) [15] includes a personal and smart DSS (Decision Support System) that can forecast a telephone call arising from long-term deposit transactions using the DM system. This research carried out and evaluated four differential classification DM models in the R tool: logistic regression (LR), decision tree (DT), neural network (NN) and support vector machine (SVM). All these techniques were measured using two parameters: the area of the receiver operating characteristic curve (AUC) and the area of the LIFT composite curve (ALIFT). The forecast of these models is carried out in a most reliable way by cross validation using metrics.

Kim, J-B et al. (2015) [16] his study analysed using a volume survey analysis of 3725 loan facilities, the impact of performing consumers on the terms and conditions of precious or non-precious loans. Three key theories were established and the findings were obtained through regression. The findings of this analysis indicate that the success of consumers makes a major difference to loan contracts.

III. Methodology

In this study, KDD-Knowledge Discovery in Databases data mining methodology is being inspired and implemented. In the framework of massive databases, the key objective of the KDD method is to extract information from records. It does this by using algorithms from data mining to determine what information is considered. The main objective of KDD methodology in data mining is to understand the data and data transformation for determining the appropriate data mining predictive models (Supervised and unsupervised data mining). The KDD model process is implemented and followed to inspire knowledge from data from massive databases. Using this acquired knowledge, data analysis and prediction of data is performed.

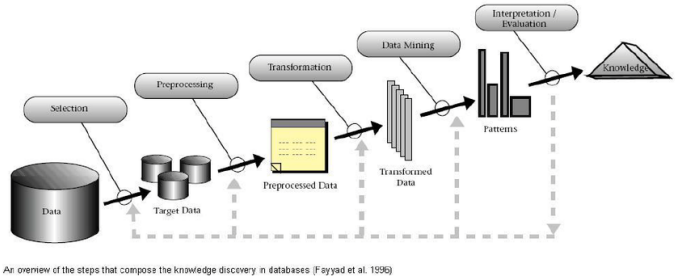

Fig. 1. KDD Methodology

Data Source

Dataset-1: Online News popularity

Dataset has 39644 records with 61 columns (58 predictive attributes, 2 non-predictive, 1 target variable). This dataset is referred from Mashable (www.mashable.com).

Link- https://data.world/uci/online-news-popularity

Dataset-2: Reused Cars Price Prediction

Dataset includes reused Craigslist car data. Craigslist is the largest collection of affordable vehicles registered in the United States. There are 68478 rows and 20 attributes.

Link-https://www.kaggle.com/austinreese/craigslist-carstrucks-data

Dataset-3: Bank Marketing Terms Client Acceptance

Dataset is linked to a Portuguese banking institution's direct marketing campaigns. Bank-additional-full.csv file has all examples with 21 attributes and 41188 records.

Link-http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

Data Cleaning and Exploratory Data Analysis

Dataset-1: Online News Popularity

The online news prediction dataset has a large volume of data (39644 rows) with 61 columns. With the help of python scripting, basic data cleaning and advanced EDA is carried out. To begin with dropping of non-predictive columns were dropped, check for NA, null values and any dirty data is carried out and cleaned. Under EDA section, the target variable "Shares" values are labelled based on the grade level based on the percentage- Exceptional = Top 95% Excellent = Top 90% Very Good = Top 80% Good = Top 60% Average = Top 50% Poor = Top 35% Very Poor = Rest and updated with a new column name "Popularity". Then similar columns were merged together for easy analysis.

Dataset-2: Reused Cars Price Prediction

The reused cars dataset has a huge volume of data (68478 rows) with 20 columns. Using the python scripting, proper data pre-processing and EDA is performed to gain a clean data for better model fitting. Non-predictive columns were dropped from the dataset. As a part of EDA, some of the categorical columns were converted into integer values, "Year" and "Odometer" columns were converted into integer datatype for better analysis. Based on the research on reused cars market, the data is restricted within a price range of 1000 to 40,000 dollars.

Dataset-3: Bank Marketing Terms Client Acceptance

This bank marketing dataset with client data has 21 attributes and 41188 records. In this dataset, we have performed all the basic data cleansing checks. We have performed an interesting EDA on this dataset by using Crosstab analysis (i.e., comparing each independent variable with the target variable "y" (yes/no)) since there were almost 20% of unknown values, comparisons were made to handle the missing values.

Data Mining Models Implemented

In this phase, selection of data mining techniques is carefully chosen according to the nature of data and prediction of each dataset. As proposed earlier, total of 5 data mining models are shortlisted such as **Linear regression, Random forest, Decision tree, K–Nearest Neighboring (KNN) and Logistic regression**. Furthermore, reason for choosing these models are explained in detail below.

Applied Data Mining Models

(1) *Random Forest:*

Random Forests (RF), consisting of a combination of tree predictors to the degree that each tree relies on the values of a separately partitioned random vector together with a related distribution across all forest trees. It uses an internal approximation feature to track inaccuracies, and to calculate the frequency of similar / dissimilar relationships [7]. For this analysis, RF uses a predictive modelling classifier to assess clients who are likely to subscribe a term deposit (Dataset 1). RF model is also implemented to predict the reused cars price (Dataset 2)

(2) *K– Nearest Neighboring (K-NN):*

K– Nearest Neighboring (KNN) is a simple, easy-to-use, monitor-controlled machine-learning algorithm, that can be utilized for solving both classification problems and regression issues. KNN is a lazy learning algorithm as it does not have a specific training phase and incorporates all the data for training while classifying. This classification model is implemented in predicting the online news popularity (Dataset 1) considered the observed values.

(3) *Linear Regression:*

The Linear regression is mostly commonly used type of predictive analysis. The main focus of regression model is to describe the relationship between the dependent variable (y) and one or more independent variables (x). Regression model also examines factors such as how good the set of predictors variables contribute for the prediction of target variable. And also determines the significance level of variables that is being used for prediction. We have implemented Linear regression model in dataset 2 for predicting the target variable "price" using multiple independent variables so it can be termed as **Multiple Linear Regression**.

Multiple linear regression model,

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i, i = 1, ..., n \quad (1)$$

(4) *Logistic Regression:*

Logistic regression (LR) has the benefit of identifying relations between nominal (binary) target variables and one or more continuous predictor variables. In fact, LR uses the highest likelihood approximation. In this process, the actual values of the expected parameters are used and the likelihood of the sample originating from a population with these parameters is estimated. The values of the predicted parameters are then modified regularly until the highest probability value is achieved [17]. The dataset- 3 consists of constant independent variables and a binary target variable, it is reasonable to use LR as a classifier to assess the customer's subscription to the term deposit of the banking telemarketing campaign under review.

(5) *Decision Tree:*

The Decision tress is a simple and easy to implement algorithm which gives the final output in graphical tree structure. It works by splitting data points into two or multiple subcategories to ensure the homogeneity of subgroups. This model iteratively divides data sets into subsections to create a tree structure to improve predict accuracy [17]. Since the dataset-3 contains the categorical data in the target variable and

helped by numerical data of independent variables, Decision tree can handle this data to produce classification predictions.

IV. RESULT INTERPRETATION AND EVALUATION

The above-mentioned data mining models has been applied on the respective datasets. We have followed hold out methodology to split the data into training and test data which is been passed as an input parameter for a model fit. Based on the train data and test data prediction will be made by the respective model fit. The model outcome and the results are being evaluated and explained in detail.

K–Nearest Neighboring (KNN) model: *Online News popularity (Dataset-1)*

KNN model is applied to predict the popularity of online news or articles. The popularity of the news or articles has been graded from poor to excellent for prediction. The model assumptions are verified for the data before fitting the model. The data has been divided into train and test data where train data holds 70% of the data and test data has 20% of the data. The required independent variables are selected for model fit. Form the sklearn,neighbors package KNeighborsClassifier has been imported to run the KNN model.

From the below result, KNN model with K=7 has been fitted. It is seen that the KNN model was able to achieve the accuracy of 44% only for the test data. The comparing metrics like Precision score, recall, f1-score were checked. As we see from the below result, we have got the precision, recall and f1- score for the nearest 7 neighbors.

Accuracy on Test Data: 0.44%

	precision	recall	f1-score	support
0	0.38	0.44	0.41	4007
1	0.00	0.00	0.00	18
2	0.00	0.00	0.00	14
3	0.52	0.61	0.56	5077
4	0.20	0.08	0.12	1560
5	0.00	0.00	0.00	77
6	0.13	0.04	0.06	786
accuracy			0.44	11539
macro avg	0.18	0.17	0.16	11539
weighted avg	0.40	0.44	0.41	11539

Fig. 2. Classification Report of KNN

Since, KNN was able to achieve only 44% of accuracy we have implemented Random forest model for the same dataset.

Comparison between KNN and Random forest model

Random Forest was able to deliver marginally higher accuracy as compared to the KNN model to obtain an accuracy of 50%. Due to the design of Random forest being able to set a specific number of decision trees, attributes, tree size, splitting parameters, and others, a lot of parameter tuning is required.

On comparing the precision, recall, f1-score of random forest model and KNN model it is seen that Random forest gives the best accuracy than KNN model.

Accuracy on Test Data: 0.50%

	precision	recall	f1-score	support
0	0.43	0.35	0.39	4007
1	0.00	0.00	0.00	18
2	0.00	0.00	0.00	14
3	0.53	0.85	0.65	5077
4	0.19	0.00	0.01	1560
5	0.00	0.00	0.00	77
6	0.19	0.01	0.01	786
accuracy			0.50	11539
macro avg	0.19	0.17	0.15	11539
weighted avg	0.42	0.50	0.42	11539

Fig. 3. Confusion Matrix of Random Forest

The best machine learning model was the Random Forest, which was able to achieve 50% accuracy on the test data collection. Some of the reasons for this low accuracy score is due to large variation in the data set and also to the imbalance in the class distribution that causes the predictive models to be skewed towards the popularity groups with more papers.

Linear Regression model: *Reused Cars Price Prediction (Dataset-2)*

Using Jupyter IDE an open-source web app data has been used for performing the respective model for reused cars price prediction dataset. Data has been imported, pre-processed and transformed before fitting in the model. We collected the data and in pre-processing step normality distribution has been checked. It is noticed that re-used car price the response variable is not normally distributed it is highly skewed. For normalizing the data, we have taken log transform to the response variable and the same gets normally distributed as shown in below figure:

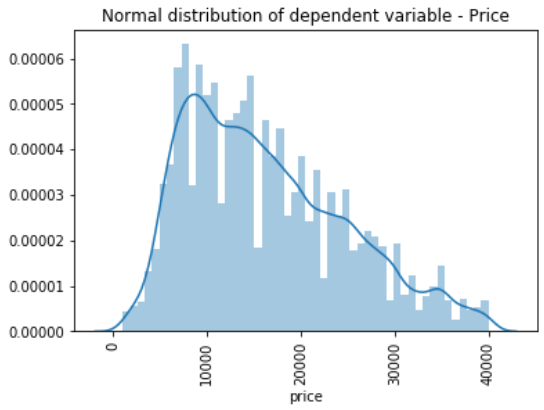

Fig. 4. Normality Distribution of Reused Car Price

After transformation of data, machine learning method has been implemented. Above mentioned two techniques are implemented for the reused car dataset. In order to determine the efficient technique for prediction, we require few parameters. Then we make comparison among Linear regression and Random Forest techniques. The parameters chosen are Accuracy

(R square value) and Root Mean Square Error [RMSE]. **R2** value for linear regression is **57.11** and RMSE is 559796.73.

```
target =  [19500 18500  7995 10900  8500]
ytrain =  [24424.78741106 17928.28221722  6571.30338779  5243.37585176
 14858.8262548 ]
acc(r2_score) for train = 57.82
acc(relative error) for train = 25.22
acc(rmse) for train = 553496.76
target_test = [ 4800 13398 20988 28900 11731]
ytest = [12536.57451931  8944.60643922 13445.69187854 23966.25124505
 10257.43180946]
acc(r2_score) for test = 57.11
acc(relative error) for test = 25.44
acc(rmse) for test = 559796.73
```

Fig. 5. Linear Regression Results

Comparision Linear regression Vs Random forest model

```
target =  [19500 18500  7995 10900  8500]
ytrain =  [19868.5   17999.5    9180.625  9659.5    8950.   ]
acc(r2_score) for train = 96.72
acc(relative error) for train = 5.41
acc(rmse) for train = 154368.48
target_test = [ 4800 13398 20988 28900 11731]
ytest = [ 3190.   8854.2 20688.7 28900.  11384.9]
acc(r2_score) for test = 84.51
acc(relative error) for test = 12.67
acc(rmse) for test = 336366.48
```

Fig. 6. Random Forest Results

From the above figures, we compared both the algorithms to find the best model. **R2** value for linear regression is **57.11** and RMSE is 559796.73. On the other hand, **R2** value for random forest is **84.51** with the RMSE value of 336366.48. It is concluded saying that random forest model performs better for the respective dataset than linear regression with the **accuracy of 84.51%**

Logistic regression model and Decision tree model: *Bank marketing terms client acceptance (Dataset-3)*

Transformed data is divided as training data and test data with the weightage of 80:20 percentage. The training and test set of data is fitted into logistic regression model and further to calculate the efficiency and performance of the model certain evaluation tests are applied.

From the below output of logistic regression, it is seen that only post come success is not statistically significant while all the other independent variables are statistically significant with p-value less than 0.05.

In order to evaluate the model performance, confusion metrics is calculated using caret package in R. The confusion matrix we can infer that true positive (TP) value = 25535 is the correctly predicted values in the model. True negative (TN) value= 2081, which is wrongly predicted values. The false positive value = 1571 which determines the miss prediction rate and false negative value = 2081. This means that client acceptance of bank term is predicted incorrectly for 2081 times. Kappa value for the model is 0.41%. The sensitivity and specificity are the important factors in evaluation of the

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1363  -0.3951  -0.3346  -0.2382   2.7988

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         65.9168072  3.8313927  17.204  < 2e-16 ***
agemid              -0.1769233  0.0487356  -3.630 0.000283 ***
agehigh              0.1785377  0.0950352   1.879 0.060293 .
contactcellular      0.4059889  0.0610789   6.647 2.99e-11 ***
month(05)may        -1.5308954  0.1143478 -13.388  < 2e-16 ***
month(06)jun        -0.4872326  0.1247922  -3.904 9.45e-05 ***
month(07)jul        -0.5316558  0.1246735  -4.264 2.00e-05 ***
month(08)aug        -0.7590794  0.1211852  -6.264 3.76e-10 ***
month(10)oct        -0.8283920  0.1417583  -5.844 5.11e-09 ***
month(11)nov        -1.0477606  0.1256554  -8.338  < 2e-16 ***
month(12)dec        -0.5671038  0.2090464  -2.713 0.006671 **
month(04)apr        -0.8061365  0.1194677  -6.748 1.50e-11 ***
month(09)sep        -1.1521548  0.1480853  -7.780 7.23e-15 ***
day_of_week(02)tue   0.2561945  0.0643412   3.982 6.84e-05 ***
day_of_week(03)wed   0.3487747  0.0640040   5.449 5.06e-08 ***
day_of_week(04)thu   0.2890088  0.0621679   4.649 3.34e-06 ***
day_of_week(05)fri   0.2376569  0.0647390   3.671 0.000242 ***
poutcomefailure     -0.5213575  0.0631761  -8.252  < 2e-16 ***
poutcomesuccess      0.2911045  0.2184294   1.333 0.182625
cons.price.idx      -0.0954891  0.0391932  -2.436 0.014836 *
nr.employed         -0.0113909  0.0003334 -34.164  < 2e-16 ***
pdays_dummy1         1.1133871  0.2065204   5.391 7.00e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22707  on 31930  degrees of freedom
Residual deviance: 17845  on 31909  degrees of freedom
AIC: 17889
```

Fig. 7. Logistic Regression - Summary of the Model

model. Sensitivity or recall value = 71% and specificity or precision value is 74%. Finally, the model as managed to get the overall accuracy of 86% which is quite satisfying.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 25535  1571
         1  2744  2081

               Accuracy : 0.8649
                 95% CI : (0.8611, 0.8686)
    No Information Rate : 0.8856
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4148

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

            Sensitivity : 0.56982
            Specificity : 0.90297
         Pos Pred Value : 0.43130
         Neg Pred Value : 0.94204
              Precision : 0.43130
                 Recall : 0.56982
                     F1 : 0.49098
             Prevalence : 0.11437
         Detection Rate : 0.06517
   Detection Prevalence : 0.15111
      Balanced Accuracy : 0.73640

       'Positive' Class : 1
```

Fig. 8. Confusion Matrix of Logistic Regression

Decision Tree

In this dataset we are trying to predict the customer term acceptance by using important bank marketing parameters. Since the dependent variable consist of categorical value (0's and 1's) where 0- term accepted and 1- term not accepted.

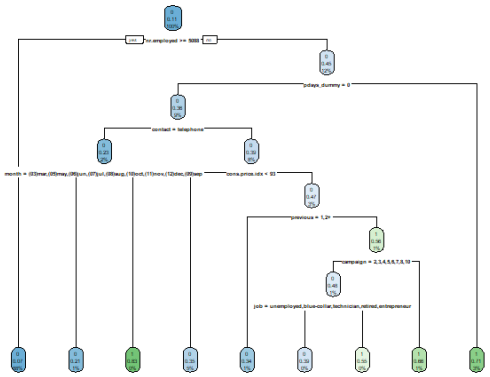

Fig. 9. Decision Tree

From the below result of decision tree. The confusion matrix is giving the result of true positive (TP) value = 26477 is the correctly predicted values in the model. True negative (TN) value= 1673, which is wrongly predicted values. The false positive value = 1979 which determines the miss prediction rate and false negative value = 1802. This means that client acceptance of bank term is predicted incorrectly for 1802 times. Kappa value for the model is 0.40%. The sensitivity and specificity are the important factors in evaluation of the model. Sensitivity or recall value = 45% and specificity or precision value is 93%. This model got the accuracy of 88% which is slightly higher than logistic regression model.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 26477  1979
         1  1802  1673

               Accuracy : 0.8816
                 95% CI : (0.878, 0.8851)
    No Information Rate : 0.8856
    P-Value [Acc > NIR] : 0.988329

                  Kappa : 0.4029

 Mcnemar's Test P-Value : 0.004206

            Sensitivity : 0.45811
            Specificity : 0.93628
         Pos Pred Value : 0.48144
         Neg Pred Value : 0.93045
              Precision : 0.48144
                 Recall : 0.45811
                     F1 : 0.46948
             Prevalence : 0.11437
         Detection Rate : 0.05239
   Detection Prevalence : 0.10883
      Balanced Accuracy : 0.69719

       'Positive' Class : 1
```

Fig. 10. Confusion Matrix of Decision Tree

V. Conclusion and Future Work

The proposed project, in three different sector which is been chosen is implemented with 5 different data mining models and their results have been evaluated. The models such as linear regression, random forest, decision tree, logistic regression and K-NN model have been used. Out of these model's logistics regression obtained 86% accuracy, decision tree obtained 88% accuracy with defines it has a best model overall. The cross validation of models has been carried out using the validation parameters to define the best fit model for each sector. In future, plan is to conduct more advanced research using deep learning techniques on these fields.

References

[1] M. T. Uddin, M. J. A. Patwary, T. Ahsan and M. S. Alam, "Predicting the popularity of online news from content metadata," 2016 International Conference on Innovations in Science, Engineering and Technology (ICISET), Dhaka, 2016, pp. 1-5.

[2] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck, "Characterizing the life cycle of online news stories using social media reactions," in Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work 38; Social Computing, ser. CSCW '14. New York, NY, USA: ACM, 2014, pp. 211–223.

[3] A. Tatar, M. de Amorim, S. Fdida, and P. Antoniadis, "A survey on predicting the popularity of web content," Journal of Internet Services and Applications, vol. 5, no. 1, 2014.

[4] Gegic, Enis, Becir Isakovic, Dino Keco, Zerina Masetic, and Jasmin Kevric. "Car price prediction using machine learning techniques." TEM Journal 8, no. 1, 2019, p.113

[5] Lessmann, Stefan, and Stefan Voß. "Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy." International Journal of Forecasting 33.4, 2017, pp. 864-877.

[6] Elsalamony, Hany A. "Bank direct marketing analysis of data mining techniques." International Journal of Computer Applications 85.7, 2014, pp.12-22.

[7] J. Asare-Frempong and M. Jayabalan, "Predicting customer response to bank direct telemarketing campaign," 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T), Kuala Lumpur, 2017, pp. 1-4.

[8] H. Chen, X. Zhong, J. Sun and J. Wang, "Online prediction algorithm of the news' popularity for wireless cellular pushing," 2015 IEEE/CIC International Conference on Communications in China (ICCC), Shenzhen, 2015, pp. 1-5.

[9] I. Arapakis, B. Cambazoglu, and M. Lalmas, "On the feasibility of predicting news popularity at cold start," in Social Informatics, ser. Lecture Notes in Computer Science, L. Aiello and D. McFarland, Eds. Springer International Publishing, 2014, vol. 8851, pp. 290–299.

[10] J. G. Lee, S. Moon, and K. Salamatian, "Modeling and predicting the popularity of online contents with cox proportional hazard regression model," Neurocomputing, vol. 76, no. 1, pp. 134–145, 2012

[11] Maddah-Ali M A, Niesen U. Fundamental limits of caching[C]// Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on. IEEE, 2013: 1077-1081.

[12] Kelwin Fernandes, Pedro Vinagre, and Paulo Cortez. "Proactive Intelligent Decision Support System for Predicting the Popularity of Online News". Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.

[13] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764.

[14] Wu, J. D., Hsu, C. C., Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications, 36(4), 7809-7817.

[15] Noor, K., Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 167(9), 27-31.16

[16] Kim, Jeong-Bon, Byron Y. Song, and Yue Zhang. "Earnings performance of major customers and bank loan contracting with suppliers." Journal of Banking Finance 59, 2015, pp. 384-398.

[17] Moro, S., Cortez, P., Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22–31. https://doi.org/10.1016/j.dss.2014.03.001