Data Mining and Machine Learning in Bank Marketing and Real Estate

Sobil Dalal
School of Computing
National College of Ireland
Dublin, Ireland
x19148496@student.ncirl.ie

***Abstract*—The objective of this study is to implement data mining and machine learning models in bank marketing and real estate sectors in the financial domain and gain insights about the factors influencing house price prediction and a customer's decision to open term deposit account with a bank as a result of a marketing campaign. Multiple regression and classification models have been implemented on three datasets and performance of each has been compared. For some techniques, stratified random sampling with a 80-20 ratio has been inspected, for others 10 fold cross validation has been used. Factors including living space, grade of house, built year, number of rooms, types of floor have been considered for house price prediction using multi-linear regression with RMSE of 0.3163 for King County, USA and RMSE of 1743 for another dataset. For bank marketing classification, to ensure balanced distribution of observations, under-sampling of biased dependent class using "ROSE::ovun.sample" has been explored. A customer's decision to open a term deposit account with the Portugal bank after a marketing campaign, is highly influenced by last contact duration, number of employees and employment variation rate as identified by JRip decision tree with an accuracy of 90.013% and AUC of 0.90.**

***Index Terms*—Multiple Regression, JRip Decision Tree, Naive Bayes, k-NN, SVM, ROSE**

I. INTRODUCTION

Data Mining (DM) and Machine Learning (ML) algorithms work on historical data that facilitates understanding of the process and provides insights on the relationship between data, accumulation of knowledge and forecasting potential outcomes through predictive analysis [1], [2]. Remarkable developments have been made using ML in several fields, but it has just penetrated in financial domains. Arithmetic average choices to assess a functional relationship on real-time financial data is a challenge for finance, risk management and pricing. By generating different models with varying dependent and multiple independent variables, DM and ML techniques work better than those conventional techniques [3].

To unearth the complexity of data in financial sector, multiple DM and ML techniques have been explored by researchers such as, usage of multi-layer neural networks to evaluate complex non-linear functions with vast volumes of available data [4], application of Deep Learning (DL) models to aid decisions in retail, finance, and financial risk management [5], development of a new hybrid model called optimal multiple kernel-support vector regression for prediction of financial data [6]. Finance is divided into several sub-domains, and this project reflects on the financial aspects of real estate and banking, specifically recognizing what are the most critical factors for forecasting house prices and what is the impact of a marketing campaign on a customer's decision to open a term deposit account with the bank.

A. House Pricing

Land value plays a key role in the redeployment of land resources and creation of successful control policies. The spatial-temporal variation of land values can be illustrated by a study of influential factors [7]. Similarly, house prices vary on a very frequent basis and it is very difficult to evaluate the right price of the property. The reason behind this is that the number of factors which influence the price of houses are high and relationship between them is complex. This problem is being faced from quite sometime and multiple studies are being undertaken to evaluate the valuation of house prices to assist the real-estate agents and sellers [8]. Further, more researches are done to work out correlation between parameters affecting price of apartments, such as apartment prices in Tehran are determined based on air quality and floor level [9]. House pricing is a traditional problem, but with time, additional factors have contributed to the existing problem and a generalized model to predict the prices and analyse the factors would be highly encouraged. To achieve this objective of predicting house pricing, two data sets have been used namely the sales of houses in King County [10] and computer generated house pricing data set [11].

B. Bank Marketing

Banks have been gathering consumer behaviour data for decades, but with new high-computation machines and ML, they have started to use the same for customer service, for example Scotia bank is using ML to boost their customer base and provides authentic, customer-centered personal service and support to enhance marketing efficiency and generate higher revenues [12]. In today's market, every bank is competitive and a critical objective for them is to utilize the data of customers to increase profit. This can be achieved by using targeted and appropriate services and promotions for potential clients. On a regular basis, banks roll out campaigns to offer new services and products to improve relationship with customers [13]. In order to make full use of the data, the campaign customer feedback and related independent parameters should be analysed. This process will further help the enterprise to make intelligent decisions for future campaigns and services. The marketing dataset of Portugal Bank is used to achieve the objective of assessing the impact of the marketing campaign on the decision of the consumer to open a term deposit account with the bank [14].

This project is, therefore, focused on 3 different datasets to train diverse machine-learning models and to test and develop them by variegated methods, mainly using the Knowledge Discovery in Databases (KDD) approach, in order to respond to the above research questions. This consists of steps, such as data collection, pre-processing and data transformation, implementation of models for data mining and model output assessment [15]. In addition, a description of the results obtained from these models and potential research to enhance the analysis and increase efficiency has been presented in subsequent sections.

II. Related Work

Following are some related works in the house pricing and bank marketing domains:

A. House Pricing

Structural attributes such as number of bedrooms, bathrooms, other rooms, property size, type of amenities are the most common predictors in the model of house pricing. Several researches have confirmed that the number of bedrooms and bathrooms, and floor size are favorably associated with the house prices [16]–[19]. On the contrary, age of the property impacts the pricing negatively [20]. A case study on public housing resale markets in Singapore shows how crucial it is to use the decision tree approach to determine significant factors that contribute in home prices and to forecast the same. The home-buyers are divided using decision trees based on type of house such as two, three, four, five room flats and executive apartments. Moreover, different predictor variables help to better predict the house prices based on these categories [21]. It should be noted that decision tree might not perform better if the number of factors in an attribute increase beyond comprehension.

House prices in Turkey are analyzed using survey data from the 2004 household budget in a research paper and it is proposed that the artificial neural network model better forecasts house prices due to the possible existence of non-linearity in the hedonic function of the regression model [22]. In another case study, house prices are measured using the artificial neural network model with a range of disparate variables such as number of bedrooms, bathrooms, garage and amenities. The model does well and overcomes the non-linear interaction between house attributes and prices, and difficulties related to data patterns. It should be remembered, however, that the technique of trial and error produces an ideal artificial neural network model. Results might not be better without this technique [23]. To forecast house prices, a combination of two algorithms is used, that is, a genetic algorithm (GA) and a support vector machine (SVM) algorithm called G-SVM. SVM does well in both classification and regression, but is expensive in terms of computation and simulation time. In contrast, GA takes less time and is thus, used to refine the parameters that are later used in SVM to forecast house prices [24]. This new type of modelling attempts to improve the modelling time but this might result in leaving out useful details while refining the data using GA. Moreover, house price valuation is performed using random forest (RF), which is also known as a special case of simple regression trees and is stronger than multiple regression and neural ML model of the network. RF is proposed to effectively handle multi-level categorical data and take care of avoiding over-fitting data that arises in multiple regression and neural networks model. Also, missing values do not influence the model and it allows non-linear relation between the dependent variable and predictors [25].

The house pricing of US market in Fairfax County, Virginia is analysed using various classification models, namely RIPPER, C4.5 Decision Tree, Naive Bayes and AdaBoost on assorted residential properties and town-houses. Comparison of the accuracy is only performed for classification models [8]. In this study, only classification technique is used. It has not incorporated some of the important quantitative continuous parameters such as size, number of rooms and bathrooms, floors, and age of the house. Many of the aforementioned models are effective in forecasting house prices accurately, but the performance of model often depends on the predictors, the form of predictors, e.g. numeric, factors. Therefore, it is assumed that the data might need some transformation before implementing regression models after evaluating the properties of the data sets for house-pricing which are both quantitative and qualitative.

B. Bank Marketing

The historic exchange and transactional features are described as the most important for choosing the most suitable customer for a marketing strategy to opt for personal loans from the time series database of retail bankers. The customers are neatly classified through random forest and deep neural network modelling techniques. The strategy involves a test procedure that is not limited to the banking industry and is applied as a framework for targeted marketing campaigns in all types of businesses. Some of the disadvantages of the analysis is that it performs only binary classification and analyzes a single product for a particular customer. With respect to technical aspects, the data by which the model is trained is just for 90 days and might be biased. It is also not evaluated commercially [26].

In an analysis, the response to an offering of banking products via telephone call was classified for a set of mixed type of banking data containing both numeric attributes like age and balance, and nominal attributes like education and marital status. Multiple ML algorithms were applied on the data sets namely, neural network (NN), logistic regression (LR), support vector machine (SVM) and decision tree (DT).The NN model scored the highest response with 75 percent accuracy preceded by the LR, SVM and DT models. DT being the worst with 67.02 percent accuracy [27]. Additional two studies were carried out in relation to this analysis, one of which indicates that the deep convolutional neural network (DCNN) model is superior to classify response with 76.70 percent. DCNN exploits relationships and hierarchical features amongst attributes [28]. Another research shows that Adaboost SVM yields better results than standard SVM [29]. These modelling techniques may be ideal for the bank marketing data set because it also contains mixed type of data attributes and needs to classify the customers by either ”yes” or ”no” as levels.

To handle the classification of imbalanced data which causes the ML algorithms to fail to achieve expected results, Synthetic Minority Oversampling Technique (SMOTE) technique has been used to balance the marketing data set before applying the Naive Bayes(NB) classification model. By doing this, there is improvement in the accuracy of the NB classifier [30]. This SMOTE technique can be useful in case of imbalanced data.

Multilayer Perceptron Neural Network (MLPNN), Random Forest (RF), Logistic Regression (LR) and Decision Tree (C4.5) models are applied to categorise the customers for bank direct marketing products. RF performed best amongst all with 87 percent accuracy [31].

In one paper, the consumer conversion is evaluated using four DM techniques: multilayer neural perception network (MPLNN), tree augmented Naïve Bayes (TAN), LR and C5.0 decision tree. Amongst all, C5.0 marginally performed better than other models. As per MPLNN, LR, C5.0 decision tree model ”duration” is the most important feature, while ”age” is more significant in the TAN model [32]. Another research reveals the efficiency of the J48 decision tree (J48-DT) and the Naïve Bayes (NB) classification model to assist with direct marketing. Classification outcomes are analyzed using a confusion matrix to test precision, specificity and sensitivity. According to the report, J48-DT performs better than NB [33].

There are a few other researches made in other sectors which are relevant for classification of marketing data. In a research, a feature vector is generated using Frequent pattern growth (FP-grow) algorithm using frequent item sets in e-commerce marketing data to classify users. The Naive Bayes Algorithm is used to incorporate clustering learning for precision marketing and to deliver customized online referral services [34]. To classify the screening of candidates in an organization for a vacancy based on attributes which are of nominal type, classification models are used, namely KNN and Naive Bayes. Candidate profile attributes are various independent factors which are used to classify the candidate selection in the organization [35].

The bank marketing data in this analysis is identical to those in this section of the marketing data, where dependent variable is of classification type and independent parameters can be converted as factors, like job, marital status, month, and education. Classification type modelling technique will be a good fit to predict the decision of customer to open term-deposit based on other parameters. For the study, multiple classification models have been generated and the accuracy, sensitivity and specificity has been compared to evaluate the best model.

III. KDD Methodology

In this project KDD methodology is followed. Total 6 different models are applied on 3 data sets which includes 2 data sets of house pricing and 1 data set of bank marketing as discussed in section I. Methodology followed in each data set is discussed below in detail:

A. House Pricing: Sales of houses in King County, USA

1) Data selection: The data set has been downloaded from kaggle website in CSV format into the local machine. This dataset has 21 613 records with 21 parameters, containing both continuous and categorical parameters about the price of house sale in King County (including Seattle), USA.

2) Data exploratory analysis, pre-processing and cleaning: Id, lat, long, and zip-code parameters have been removed as these did not help in improving the model. From Fig. 1, it can be inferred that the house prices are not normally distributed and are rightly skewed.

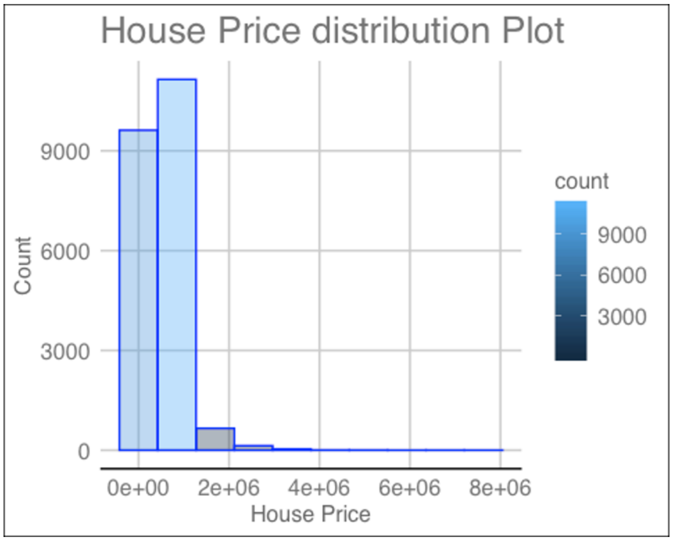


Fig. 1. Histogram: House price(dependent variable)

No NA’s are present in the data (see Appendix 1, pp. 6-7). To avoid multi-collinearity, sqft_living15 and sqft_lot15 parameters have been dropped as theses are related to sqft_living and sqft_lot, respectively. Additionally, the outliers have been removed which could have biased the modelling results and the labels of rows have been readjusted with a sequence starting from 1.

3) Data transformation and preparation: Data transformation and data preparation techniques are listed below with respect to the data mining models:

- *Multi-linear regression*: From Fig. 1 and linear model diagnostic plots of price prediction (see Appendix 1, p. 21), without any transformation in the dependent variable (price), it is evident that the dependent variable is not meeting the assumption of normal distribution. To overcome this problem, log of the price variable has been taken, which transforms it into a normal distribution (see Fig. 2).

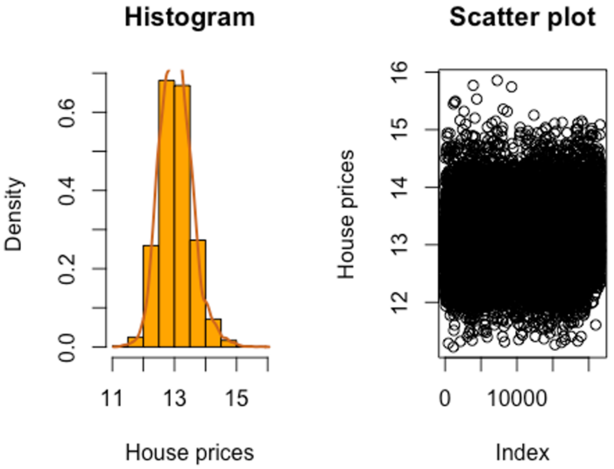


Fig. 2. Normal distribution: log(House price)

K-Fold cross validation (CV) technique has been used to evaluate the model where k is equal to 10 and thus, no training and testing samples were generated.

- *Decision tree regression*: No data transformation technique is required in this type of modelling as it combines the strength of a decision tree with the ability to model numeric data. For evaluating regression type decision tree model, stratified random sampling is used from caret package to create groups with equal representation of data. The ratio of training and testing has been set to 80-20% respectively.

4) Data mining model: The primary models on which the sample has been fitted are described below:

- *Multi-linear regression*: Function glm has been used to generate the primary linear model in which all the parameters are added in an additive manner (see Appendix 1, p. 24).
- *Decision tree regression*: Regression tree model is applied on the training sample using rpart::rpart function (see Appendix 1, p. 76).

5) Model Evaluation and Interpretation: The primary model summary is explained below:

- *Multi-linear regression*: 10-fold CV technique has been used to evaluate the performance of the model using ”boot::cv.glm” function. AIC value for the primary model has been reported as 11147 with residual deviance of 2116.7. In the model summary, the p-value for sqft_lot, sqft_basement, and yr_renovated are greater than 0.05 and thus, suggests that these predictors are not significant for predicting house price (see Appendix 1, p. 25).
- *Decision tree regression*: The root mean squared error (RMSE) of the regression model has been reported as 0.3780299 (see Appendix 1, p. 82). Furthermore, the decision tree to predict the house price is generated using rpart.plot function and is shown in Fig. 3. It can be clearly interpreted from the Fig. 3 that grade, sqft_living, and yr_built are the most important predictors to evaluate the house price.

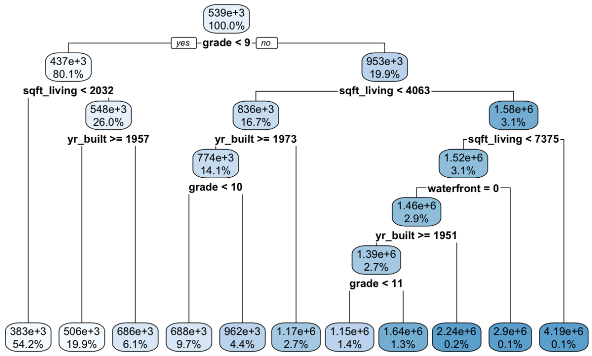


Fig. 3. Regression Tree: Visualization of decision to predict price

6) Improving model performance: In both the models multiple steps have been done to improve the performance and are listed in the following subsections:

- *Multi-linear regression*: Firstly, all the non-significant variables mentioned in above steps were removed one by one using ”update” function and the models were re-evaluated. Overall, the performance of the model increased. Secondly, all the related independent variables have been added to model as interaction terms which were identified from psych::pair.panels function plot. Thirdly, using manual step backward technique, non-significant variables have been removed again using update function (total 15 model updations). Finally, using regsubsets function the best predictors have been identified for different R-squared values (see Fig. 4). By

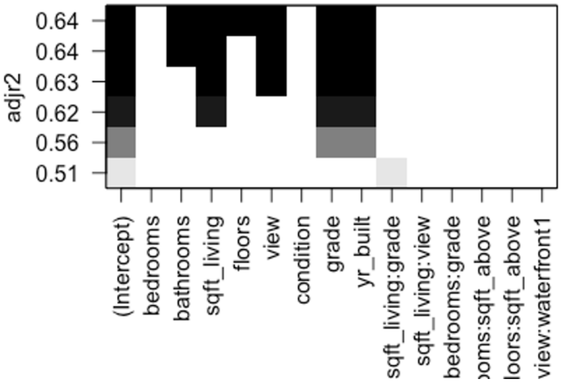
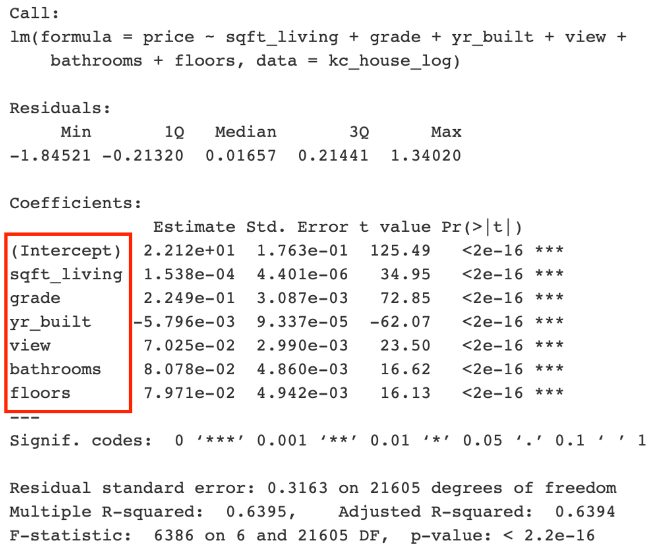


Fig. 4. Regsubplot: best predictors to fit the model

following the concept of parsimony and iterative steps to improve the performance ”house.k10.6ii” model has been generated with AIC of 11587, CV delta value of 0.1001, adjusted R-squared value of 0.6394 and RMSE of 0.3163 (see Appendix 1, pp. 24-46). Fig. 5 shows the summary and Fig. 6 shows the model performance visually.



```
Call:
lm(formula = price ~ sqft_living + grade + yr_built + view +
    bathrooms + floors, data = kc_house_log)

Residuals:
     Min       1Q   Median       3Q      Max
-1.84521 -0.21320  0.01657  0.21441  1.34020

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.212e+01  1.763e-01  125.49   <2e-16 ***
sqft_living  1.538e-04  4.401e-06   34.95   <2e-16 ***
grade        2.249e-01  3.087e-03   72.85   <2e-16 ***
yr_built    -5.796e-03  9.337e-05  -62.07   <2e-16 ***
view         7.025e-02  2.990e-03   23.50   <2e-16 ***
bathrooms    8.078e-02  4.860e-03   16.62   <2e-16 ***
floors       7.971e-02  4.942e-03   16.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3163 on 21605 degrees of freedom
Multiple R-squared:  0.6395,    Adjusted R-squared:  0.6394
F-statistic:  6386 on 6 and 21605 DF,  p-value: < 2.2e-16
```

Fig. 5. Multi-linear model summary

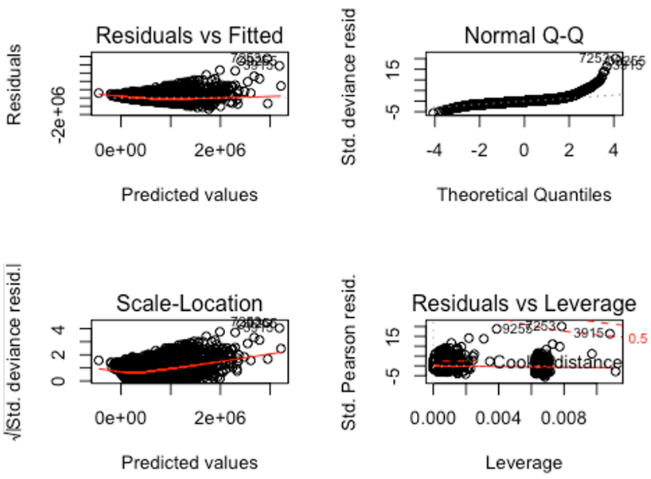


Fig. 6. Multi-linear model plot

Additionally, another two sets of linear regression modelling have been performed which contains all the steps of KDD methodology. In one type of modelling, the normal distribution assumption is ignored and the best model has adjusted R-squared value of 0.5782 (see Appendix 1, pp. 7-21). Whereas, in the other model, only numeric continuous and highly correlated data has been used to predict house price. Similarly, the best model in this case has adjusted R-squared value of 0.5646, CV delta value of 0.1208 and AIC of 15644 (see Appendix 1, pp. 46-74).

- *Decision tree regression*: Model tree has been applied to the same testing sample using RWeka::M5P function. The RMSE for this model increased to 1.794102 (see Appendix 1, p. 82).

B. House Pricing: Computer generated house pricing

1) Data selection: This data set has also been downloaded from kaggle website in CSV format. This dataset is computer generated and contains 16 continuous parameters of the house including price. It has about half a million records.

2) Data exploratory analysis, pre-processing and cleaning: As the files contains half million records ”fread” function is used to load data in data table to improve the overall performance when compared to data frames. All the attributes in the dataset have numeric continuous values. No NA’s have been found in the dataset. Furthermore, the outliers in Prices variable have been removed by identifying them using boxplot$out function (see Appendix 2, pp. 83-87). The distribution of the dependent variable Prices is normal (see Fig. 7).

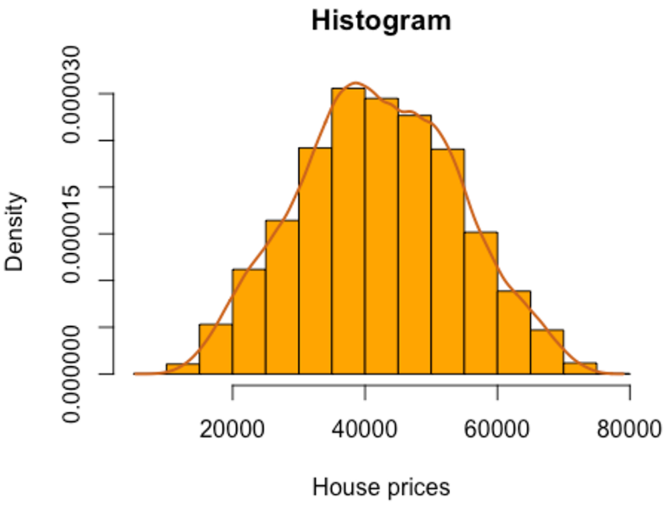


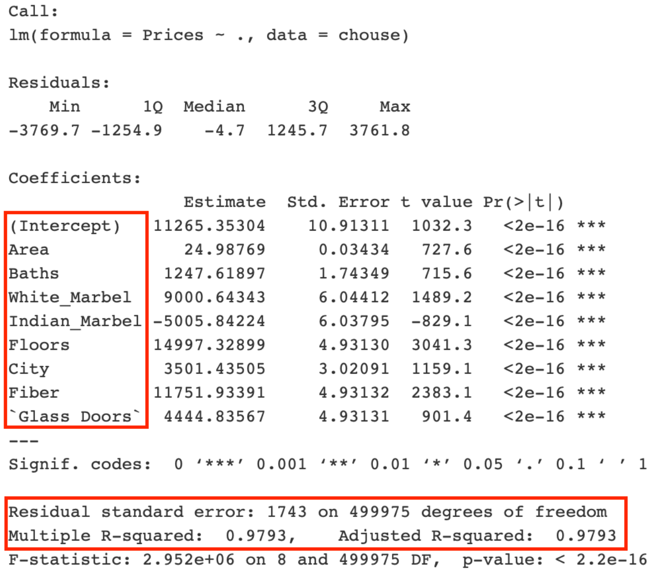
Fig. 7. Multi-linear model plot

3) Data transformation and preparation: Using the correlation matrix, a new data table is created containing attributes having correlation coefficients greater than 0.2. Also, a few columns have been renamed for better readability (see Appendix 2, pp. 83-87).

4) Data mining model: The glm function has been used to generate a primary *multi-linear model* with additive inclusion of all of the parameters and the interaction term which has been identified from the correlation matrix as ”White_Marbel:Indian_Marbel” (see Appendix 1, pp. 90-91).

5) Model Evaluation and Interpretation: The 10-fold CV technique has been used to assess the model’s performance with the ”boot::cv.glm” function. AIC was estimated as 8882283 for the primary model. Additionally, the interaction term has been found to be non-significant (see Appendix 2, p. 91).

6) Improving model performance: Initially, the interaction term which has been found insignificant was removed using "update" function and for this model the AIC remained same. The adjusted R-squared value of the model is 0.9793 and RMSE is 1743. Fig. 8 shows the summary and 9 shows the model performance.



```
Call:
lm(formula = Prices ~ ., data = chouse)

Residuals:
    Min      1Q  Median      3Q     Max 
-3769.7 -1254.9    -4.7  1245.7  3761.8 

Coefficients:
                 Estimate  Std. Error t value Pr(>|t|)    
(Intercept)   11265.35304    10.91311  1032.3   <2e-16 ***
Area             24.98769     0.03434   727.6   <2e-16 ***
Baths          1247.61897     1.74349   715.6   <2e-16 ***
White_Marbel   9000.64343     6.04412  1489.2   <2e-16 ***
Indian_Marbel -5005.84224     6.03795  -829.1   <2e-16 ***
Floors        14997.32899     4.93130  3041.3   <2e-16 ***
City           3501.43505     3.02091  1159.1   <2e-16 ***
Fiber         11751.93391     4.93132  2383.1   <2e-16 ***
`Glass Doors`  4444.83567     4.93131   901.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1743 on 499975 degrees of freedom
Multiple R-squared:  0.9793,	Adjusted R-squared:  0.9793 
F-statistic: 2.952e+06 on 8 and 499975 DF,  p-value: < 2.2e-16
```

Fig. 8. Multi-linear model summary

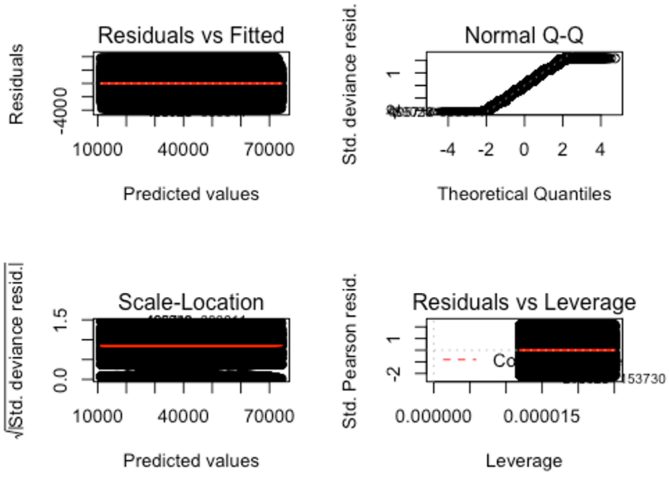


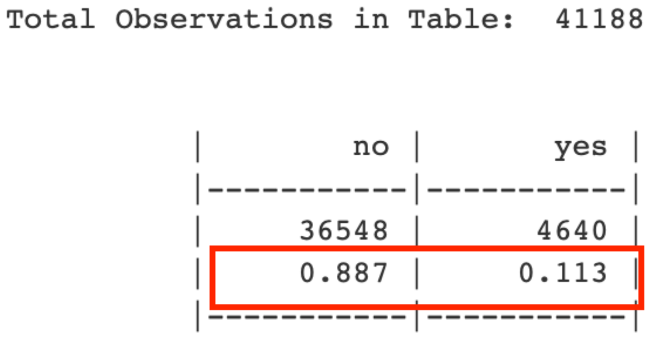
Fig. 9. Multi-linear model plot

Later, the best predictors have been chosen using "regsubsets" function having "nvmax" equals to 6. Multiple models have been generated with 4, 5, 6 best predictors which generated an adjusted R-sqaured value of 0.8742, 0.9077 and 0.9361 respectively (see Appendix 2, pp. 91-99).

C. Bank Marketing: Portugal bank marketing campaign

1) Data selection: The data has been downloaded from University of California, Irvine machine learning repository website in CSV format into the local machine. This dataset is about the decision (yes/no) of clients to open a term-deposit account in Portugal bank. It has 17 parameters and 45 211 rows.

2) Data exploratory analysis, pre-processing and cleaning: The dataset has been read into a dataframe which contains both continuous and categorical data. The dependent variable "y" has categorical data with levels "yes" and "no" in 0.113 to 0.887 ratio, respectively (see Fig. 10). No NA's have been found in the data (see Appendix 3, p. 104).



```
Total Observations in Table:  41188 


          |        no |       yes | 
          |-----------|-----------|
          |     36548 |      4640 | 
          |     0.887 |     0.113 | 
          |-----------|-----------|
```

Fig. 10. Dependent variable distribution

3) Data transformation and preparation: Different data transformation techniques have been applied with respect to the machine learning model. Each of them are listed below:

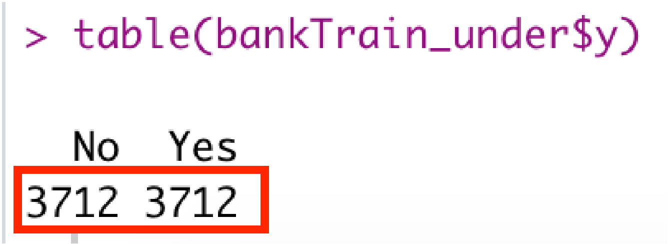
- *KNN model*: Three types of transformation have generated three different types of sample data to get the best performance out of KNN modelling. All are listed below and their performance is discussed in later sections:
    1) A custom function has been created to normalize the data using minimum and maximum values. This method has been applied to all the continuous variables in the data frame. For all the remaining categorical attributes, dummy columns have been created using "fastDummies::dummy_cols" function as KNN works well with only numeric values (see Appendix 3, p. 104-106).
    2) Inbuilt base r function "scale" has been used to normalise the continuous variable by using z-score standardisation. Additionally, dummy variables have been created as previously mentioned (see Appendix 3, pp. 120-123).
    3) Only numeric continuous normalized attributes have been added to new data frame. In this case, again "scale" function has been used (see Appendix 3, pp. 135-136).
- *Naive Bayes*: As Naive Bayes methodology only works with categorical data, all the numeric continuous variables have been converted to factors with levels. Age, duration, campaign, pdays, cons.price.idx, cons.conf.idx, euribor3m attributes have been explored using histogram binning and converted to factors using "cut" function. Rest of the continuous attributes has been directly converted to factors using "as.factor" function, as these

attributes only have a few levels which have been verified by using ”table” function. A few NA’s which appeared after conversion have been omitted, as the sample size is quite big (see Appendix 3, pp. 149-157).

- *Decision tree*: No data transformation or cleaning has been performed for decision tree, as it can handle both nominal and numeric features.
- *SVM*: Again for SVM model no data transformation is done, as it can also handle both nominal and numeric features. Apart from that, it works well with noisy data.

Using the ”caret::createDataPartition” function, stratified random sampling has been used to create two samples: training and testing sample in 80-20% ratio, such that both the samples have equal categorical values of ”yes” and ”no” classes. One instance of this can be seen in Appendix 3, p. 106.

In addition to all other steps, under-sampling has been carried out for handling imbalance in dependent variable. Over-sampling of data was also tried on the data, but due to full-heap issues faced in over-sampling, only under-sampling has been used to generate models using ”ROSE::ovun.sample” function. The distribution of dependent variable can be seen in Fig. 11. This under-sampled data set has only been used in the case of best model amongst all models.



```
> table(bankTrain_under$y)

  No  Yes
3712 3712
```

Fig. 11. Under sample training data distribution

4) Data mining model: The primary models on which the training sample has been fitted are described below:

- *KNN model*: Using the ”class::knn” function, a primary KNN model has been applied to custom normalised data set as mentioned in above subsection with k value equal to 203, which is approximately the closet odd square root value of total observations (see Appendix 3, p. 106).
- *Naive Bayes*: For training the Naive Bayes model ”e1071::naiveBayes” function has been used (see Appendix 3, p. 160).
- *Decision tree*: A primary model has been fit on the training sample using ”C50::C5.0” function, as it performs really well (see Appendix 3, p. 164).
- *SVM*: Using ”kernlab::ksvm” function, a primary model has been fit on the training data set. Initially, simple linear hyperplane is considered and therefore, ”vanilladot” is passed as argument in the kernel parameter (see Appendix 3, p. 175).

5) Model Evaluation and Interpretation: Independent model evaluation and interpretation is described below. In each model, the evaluation has been done based on the results produced by ”caret::confusionMatrix” function.

- *KNN model*: The model performed very poorly in classifying ”yes” class in the dependent variable with kappa value of 0.2307 and sensitivity of just 0.15409, even though the accuracy is 90.04% (see Appendix 3, pp. 106-107).
- *Naive Bayes*: The model performed fairly well in terms of classifying both the classes of dependent variable. Overall, the model accuracy is 88.08% and, kappa, sensitivity and specificity are 0.4313, 0.59315 and 0.91250 respectively (see Appendix 3, p. 160).
- *Decision tree*: The performance of the model is adequate in terms of classifying the dependent variable. The accuracy, kappa, sensitivity and specificity of the model are 91.85%, 0.5623, 0.55927 and 0.96415 respectively (see Appendix 3, pp. 164-165).
- *SVM*: The primary model poorly classified the ”yes” class of dependent variable with accuracy, kappa, sensitivity and specificity of 90.79%, 0.4051, 0.33297 and 0.98085 respectively (see Appendix 3, pp. 176).

6) Improving model performance:

- *KNN model*: Firstly, for the same normalised sample data, multiple k values ranging from 121 to 1 have been used to fine tune the performance of the model. For custom normalised sample, the best results have been produced using 3 as K value. The values for this enhanced model include 88.99%, 0.3282, 0.30496 and 0.96415 for accuracy, kappa, sensitivity and specificity values, respectively (see Appendix 3, pp. 107-120).
  Secondly, the ”scale” function data sample has been used to train the model with the aforementioned k values. The best results for this sample have been achieved by using k value of 9. The values for this enhanced model include 90.87%, 0.4736, 0.44289 and 0.96785 for accuracy, kappa, sensitivity and specificity values, respectively (see Appendix 3, pp. 124-134).
  Thirdly, only the scaled numeric data sample has been used to train the model with similar k values. K value of 19 has produced the best results with 91.67%, 0.5324, 0.50539, 0.96894, 0.67482, 0.50539 and 0.57794 for accuracy, kappa, sensitivity, specificity, precision, recall and F1 values, respectively (see Appendix 3, pp. 137-146).
- *Naive Bayes*: To improve the model performance, Laplace estimator value as 1 has been used to decrease the number of false positives. The new model’s accuracy, kappa, sensitivity, specificity, precision, recall and F1 values have been reported as 88.17%, 0.4335, 0.59315, 0.91344, 0.43012, 0.59315 and 0.49865 respectively (see Appendix 3, pp. 160-162).
- *Decision tree*: Firstly, to improve the performance of the model, boosting has been used with trial values of 5 and 10. It did not improve the overall performance much. Secondly, a cost matrix has been created where true positives costed 4 times than false negatives and has been passed to costs parameter of ”C50::C5.0”. This also did

not improve the performance of the model. Thirdly, 1R, a ripper algorithm model has been applied to the training data sample using ”RWeka” package ”OneR” and ”JRip” function. ”JRip” model’s accuracy, kappa, sensitivity, specificity, precision, recall and F1 values have been recorded as 91.68%, 0.5667, 0.58513, 0.95895, 0.64413, 0.58513 and 0.61321, respectively (see Appendix 3, pp. 165-173).

- *SVM*: Initially, different kernels have been tried to improve the model’s performance, namely ”rbfdot”, ”polydot”, ”tanhdot”. Amongst all, ”rbfdot” performed better. Therefore, using the kernel ”rbfdot”, cost parameter having values like 10, 50 and 100 has been introduced to different models. To sum up, the model with cost equal to 100 performed best with accuracy, kappa, sensitivity, specificity, precision, recall and F1 values as 90.14% , 0.4857, 0.51509, 0.95047, 0.56905, 0.51509 and 0.54072, respectively (see Appendix 3, pp. 177-181).

In addition to all other models, under-sampled training data sample has also been fitted on ”JRip” decision tree model. The reason and results of the same are discussed in section IV.

IV. OVERALL EVALUATION AND KNOWLEDGE DISCOVERY

In this section, evaluation of independent models and comparison of model performances are explained for each data set. Finally, knowledge discovery is also highlighted.

A. House Pricing: Sales of houses in King County, USA

In case of multi-linear model, as mentioned in previous section 10-Fold CV has been used and the best performance is achieved by log transformed price as dependent variable having adjusted R-squared value of 0.6394 and RMSE of 0.3163, when compared to adjusted R-squared value of 0.5782 for non-transformed dependent variable and 0.5646 value for only numeric and highly correlated predictors. Furthermore, in Fig. 6, from residual vs fitted plot, it can be inferred that predictors and dependent variables have correct linear functional form and do not have a pattern. From scale-location plot, assumption of homoscedasticity and constant variance of errors are verified as, most residual data points are clustered at 0 in scatter-plot and have a relatively rectangular form. Additionally, in Q-Q plot the residuals are fairly normally distributed and in leverage plot, no influential data points can be seen and therefore, validates the assumption of normal distribution of errors. In case of regression decision tree, regression tree has performed better over model tree having 0.3780299 RMSE value when compared to 1.794102. These results have been achieved using 80-20% ratio between training and testing, respectively through stratified random sampling.

Amongst multi-linear and regression tree models, the former performed better as multi-linear has RMSE of 0.3163 when compared to 0.3780299 for regression model. Overall the knowledge gained from this analysis is that the most important factors to predict house prices are living space (sqft_living), grade of the house (grade), built year (yr_built), view, number of bathrooms(bathrooms) and number of floors in the house (floors).

B. House Pricing: Computer generated house pricing

Amongst all the varied models, with respect to adjusted R-squared values, RMSE and AIC, the first model after improvement that is ”house.fit2” performed best as it has highest adjusted R-squared value of 0.9793, and lowest RMSE and AIC values 1743 and 8882283, respectively. Also from Fig. 9, it is clear from the residual vs fitted plot that predictors and dependent variables have a right linear functional form and have no pattern. The assumption of homoscedasticity and continuous variance of errors have been verified from scale location plots as all the residual points are centred towards middle and do not have any funnel like shape. The cook’s distance is less than 1. Thus, there are no influential or leverage points in the model. At last, the Q-Q plot also shows that, residuals are almost normally distributed. Therefore, the model has performed fairly well to predict the house prices.

In reference to Fig. 9, it has further augmented the finding of knowledge that for predicting the house prices; area, number of bathrooms, types of floor like white marble or Indian marbel, number of floors, city, fiber and having glass doors are the most important attributes.

C. Bank Marketing: Portugal bank marketing campaign

In this analysis, recall/sensitivity is more important to identify all the target customers and false positives are more acceptable than false negatives. Having identified that, the following overall evaluation and comparison of models is done:

On comparing knn models, amongst all the different samples created using various transformation techniques discussed in previous section, only numeric continuous variables which have been normalised using inbuilt base package r ”scale” function have generated the best results. Therefore, 19 closest data points have been used by the model to classify the testing data points. Accuracy, kappa, sensitivity, specificity, precision, recall and F1 values have been generated as 91.67%, 0.5324, 0.50539, 0.96894, 0.67482, 0.50539 and 0.57794, respectively. Amongst the Naive Bayes models, the model with laplace value 1 performed slightly better than primary model with slightly increased accuracy and specificity. The model’s accuracy, kappa, sensitivity, specificity, precision, recall and F1 values have been recorded as 88.17%, 0.4335, 0.59315, 0.91344, 0.43012, 0.59315 and 0.49865, respectively. In decision tree models, the model fitted using ”RWeka::JRip” function performed best with accuracy, kappa, sensitivity, specificity, precision, recall and F1 values as 91.68%, 0.5667, 0.58513, 0.95895, 0.64413, 0.58513 and 0.61321, respectively. Overall, this model is a balanced model which has been able to classify both classes of dependent variable fairly well. In all the models created using SVM methodology, the model with kernel ”rbfdot” and cost parameter 100 performed best with accuracy, kappa, sensitivity, specificity, precision, recall and F1 values as 90.14% , 0.4857, 0.51509, 0.95047, 0.56905, 0.51509 and 0.54072, respectively.

Individually, each type of machine learning algorithm has performed moderately well. Nevertheless, Naive Bayes with laplace 1 and ”JRip” ripper decision tree models has performed comparatively better than others. Amongst the two, decision tree has an overall edge at classifying the dependent variable ”y” that is the customer decision to open the term deposit account. The sensitivity of Naive Bayes model is slightly (0.000802) greater than decision tree model, but F1 value, accuracy, kappa, sensitivity, specificity, precision and recall are all greater for JRip ripper decision tree model.

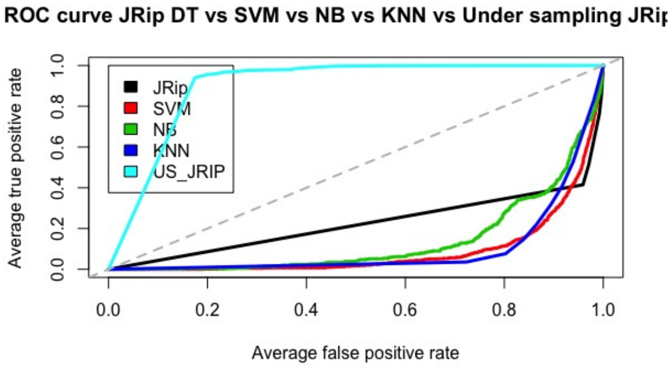


Fig. 12. ROC curve diagram

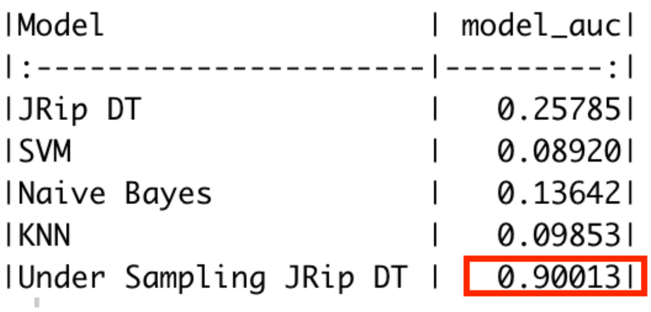

```
|Model                      | model_auc|
|:--------------------------|---------:|
|JRip DT                    |   0.25785|
|SVM                        |   0.08920|
|Naive Bayes                |   0.13642|
|KNN                        |   0.09853|
|Under Sampling JRip DT     |   0.90013|
```

Fig. 13. Comparison of models

Therefore, ”JRip” ripper decision tree has been fitted again on under-sampled training data sample. This improved the model performance drastically. The accuracy, kappa, sensitivity, specificity, precision, recall and F1 values have been evaluated as 84.19% , 0.4953, 0.9472, 0.8286, 0.4123, 0.9472 and 0.5745, respectively. Even though the accuracy, kappa, specificity and F1 values have decreased, still the sensitivity has increased tremendously. In addition to that, from Fig. 12 it can inferred that, the under-sampled JRip ripper model is very likely to identify positive values. It is almost equivalent to an ideal ROC curve. Quantitatively, it has also been verified using area under the curve (AUC). AUC has also increased by almost 7 times that is approximately from 0.137 to 0.90 (see Fig. 13). Also, from Fig. 14 it can be inferred that last contact duration (duration), number of employees (nr.employed) and employment variation rate (emp.var.rate) are the most important factors in influencing a customer’s decision to open a term deposit account with the bank.

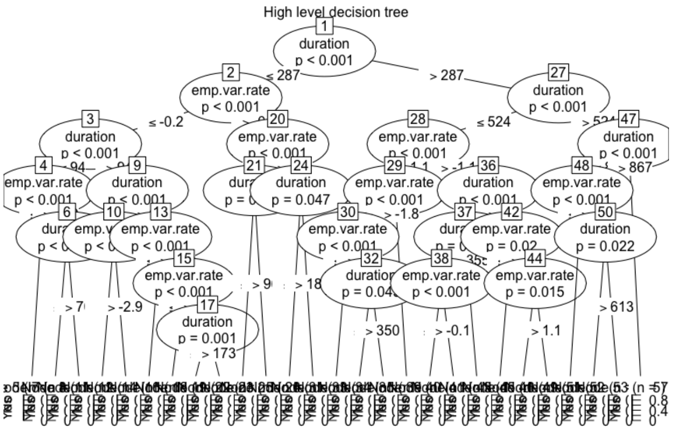


Fig. 14. High level decision tree visualisation

V. CONCLUSIONS AND FUTURE WORK

To conclude, house pricing and bank marketing sectors in the financial domain, have been explored in this project with the help of three datasets and six machine learning algorithms. For ease of analysis and balanced distribution of observations, stratified random sampling with a 80-20 ratio has been inspected for some samples along with 10 fold cross validation in others. Firstly, the most important factors influencing house price prediction have been identified as living space (sqftliving), grade of the house (grade), built year (yrbuilt), view, number of bathrooms(bathrooms) and number of floors in the house(floors) for King County, USA by training a multi-linear regression model with RMSE value 0.3163. However, for computer-generated dataset, novel factors such as type of floor (white marbel or Indian marbel), city, fiber and glass doors were also recognised by the multi-linear model with RMSE value 1743. Secondly, a customer’s decision to open a term deposit account with the Portugal bank after a marketing campaign, is highly influenced by last contact duration (duration), number of employees (nr.employed) and employment variation rate (emp.var.rate) as suggested by JRip decision tree’s results with an accuracy of 90.013% and AUC of 0.90, after under-sampling (using ”ROSE::ovun.sample”)the data to reduce bias in originally sampled data with around 11.3% values as Yes and 88.7% values as No in the dependent variable. Moreover, the house price might be affected by various other economic factors like exchange rate and interest rate that are not included in these datasets and also, development activities proposed around it. Similarly, different marketing techniques and mediums could be added to capture feedback from the customer about their willingness or concerns about opening a new term deposit with the bank. Thus, additional information with balanced samples would certainly augment classification and prediction of house prices and customer's decisions in the future.

References

[1] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining Practical Machine Learning Tools and Techniques*, 4th ed. Morgan Kaufman, 2017.

[2] J. Kelleher, B. Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, ser. The MIT Press. MIT Press, 2015. [Online]. Available: https://books.google.ie/books?id=3EtQCgAAQBAJ

[3] L. Gan, H. Wang, and Z. Yang, "Machine learning solutions to challenges in finance: An application to the pricing of financial products," *Technological Forecasting and Social Change*, vol. 153, 2020. [Online]. Available: www.scopus.com

[4] J. Cao, J. Chen, and J. C. Hull, "A neural network approach to understanding implied volatility movements," *Available at SSRN 3288067*, 2019.

[5] A. Kim, Y. Yang, S. Lessmann, T. Ma, M. . Sung, and J. E. V. Johnson, "Can deep learning predict risky retail investors? a case study in financial risk behavior forecasting," *European Journal of Operational Research*, vol. 283, no. 1, pp. 217–234, 2020, cited By :1. [Online]. Available: www.scopus.com

[6] D. Simian, F. Stoica, and A. Bărbulescu, "Automatic optimized support vector regression for financial data prediction," *Neural Computing and Applications*, vol. 32, no. 7, pp. 2383–2396, 2020. [Online]. Available: www.scopus.com

[7] J. Ma, J. C. P. Cheng, F. Jiang, W. Chen, and J. Zhang, "Analyzing driving factors of land values in urban scale based on big data and non-linear machine learning techniques," *Land Use Policy*, vol. 94, 2020, cited By :1. [Online]. Available: www.scopus.com

[8] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data." *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928 – 2934, 2015.

[9] R. Annamoradnejad, I. Annamoradnejad, T. Safarrad, and J. Habibi, "Using web mining in the analysis of housing prices: A case study of tehran," 2019, pp. 55–60, cited By 0.

[10] Harlfoxem. (2016) House sales in king county, usa. [Online]. Available: https://www.kaggle.com/harlfoxem/housesalesprediction

[11] A. Sleem. (2018) House pricing. [Online]. Available: https://www.kaggle.com/greenwing1985/housepricing

[12] Z. Abbas, R. Merbis, and A. Motruk, "Leveraging machine learning to deepen customer insight," *Applied Marketing Analytics*, vol. 5, no. 4, pp. 304–311, 2020. [Online]. Available: www.scopus.com

[13] T. Mohammad Reza, B. Seyed Mojtaba Hosseini, and T. Samrand, "A data mining method for service marketing: A case study of banking industry." *Management Science Letters*, no. 3, p. 253, 2011.

[14] S. Moro, P. Cortez, and P. Rita. (2014) A data-driven approach to predict the success of bank telemarketing. Decision Support Systems,. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/bank marketing

[15] A. Azevedo and M. F. dos Santos, "Kdd, semma and crisp-dm: a parallel overview," in *IADIS European Conf. Data Mining*, 2008.

[16] M. Fletcher, P. Gallimore, and J. Mangan, "Heteroscedasticity in hedonic house price models," *Journal of Property Research*, vol. 17, no. 2, pp. 93–108, 2000. [Online]. Available: https://doi.org/10.1080/095999100367930

[17] M. M. Li and H. J. Brown, "Micro-neighborhood externalities and hedonic housing prices," *Land Economics*, vol. 56, no. 2, pp. 125–141, 1980. [Online]. Available: http://www.jstor.org/stable/3145857

[18] G. Garrod and K. Willis, "Valuing goods' characteristics: An application of the hedonic price method to environmental attributes," *Journal of Environmental Management*, vol. 34, no. 1, pp. 59 – 76, 1992. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301479705801100

[19] M. Rodriguez and C. Sirmans, "Quantifying the value of a view in single-family housing markets," vol. 62, pp. 600–603, 1994.

[20] J. F. Kain and J. M. Quigley, "Measuring the value of housing quality," *Journal of the American Statistical Association*, vol. 65, no. 330, pp. 532–548, 1970. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1970.10481102

[21] G.-Z. Fan, S. E. Ong, and H. C. Koh, "Determinants of house price: A decision tree approach," *Urban Studies*, vol. 43, no. 12, pp. 2301–2315, 2006. [Online]. Available: https://doi.org/10.1080/00420980600990928

[22] H. Selim, "Determinants of house prices in turkey: Hedonic regression versus artificial neural network," *Expert Systems with Applications*, vol. 36, no. 2, Part 2, pp. 2843 – 2852, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417408000596

[23] V. Limsombunchai, "House price prediction: Hedonic price model vs. artificial neural network," *American Journal of Applied Sciences*, vol. 1, pp. 193–201, 2004.

[24] J. Gu, M. Zhu, and L. Jiang, "Housing price forecasting based on genetic algorithm and support vector machine," *Expert Systems with Applications*, vol. 38, no. 4, pp. 3383 – 3386, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417410009310

[25] J. Hong, H. Choi, and W. . Kim, "A house price valuation based on the random forest approach: The mass appraisal of residential property in south korea," *International Journal of Strategic Property Management*, vol. 24, no. 3, pp. 140–152, 2020. [Online]. Available: www.scopus.com

[26] P. Ładyżyński, K. Żbikowski, and P. Gawrysiak, "Direct marketing campaigns in retail banking with the use of deep learning and random forests," *Expert Systems with Applications*, vol. 134, pp. 28 – 35, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417419303471

[27] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, vol. 62, pp. 22 – 31, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016792361400061X

[28] K. Kim, C. Lee, S. Jo, and S. Cho, "Predicting the success of bank telemarketing using deep convolutional neural network," in *2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, 2015, pp. 314–317.

[29] A. Lawi, A. A. Velayaty, and Z. Zainuddin, "On identifying potential direct marketing consumers using adaptive boosted support vector machine," in *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*, 2017, pp. 1–4.

[30] M. S. Islam, M. Arifuzzaman, and M. S. Islam, "Smote approach for predicting the success of bank telemarketing," in *2019 4th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)*, 2019, pp. 1–5.

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

[32] H. A. Elsalamony, "Bank direct marketing analysis of data mining techniques," *International Journal of Computer Applications*, vol. 85, no. 7, pp. 12–22, 2014.

[33] A. A. Amponsah and K. A. Pabbi, "Enhancing direct marketing using data mining: A case of yaa asantewaa rural bank ltd. in ghana," *International Journal of Computer Applications*, vol. 153, pp. 6–12, 2016.

[34] H. Rao, Z. Zeng, and A. Liu, "Research on personalized referral service and big data mining for e-commerce with machine learning," in *2018 4th International Conference on Computer and Technology Applications (ICCTA)*, 2018, pp. 35–38.

[35] S. A. Hudli, A. V. Hudli, and A. A. Hudli, "Application of data mining to candidate screening," in *2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, Aug 2012, pp. 287–290.