Data Mining and Machine Learning in Bank Marketing and House pricing

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

***Abstract*—This research will evaluate the house pricing and bank marketing data to predict customer services of previous records. Multiple regression and classification models will be implemented on datasets and performance of each will be compared. R language will be used extensively.**

***Index Terms*—Regression modeling, Classification modeling, R language**

I. MOTIVATION

Data Mining (DM) and Machine Learning (ML) helps to understand the process and insights about the relationship between the data to accumulate knowledge [1]. To use data mining concepts in real-time, we use data science methodologies in which data mining is part of the whole process [2]. To understand the factors which affects the price of house and to determine the affects of marketing in banking, concepts of DM and ML would help analyse the same.

House prices varies on very frequent basis and it is very difficult to evaluate the right prices of the property. The reason behind this is 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 made to evaluate the valuation of house price to assist the real-estate agent and seller [3]. 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 [4]. House pricing is a traditional problem but with time, additional factors have contributed to existing problem and a generalized model to predict the price and analyse the factors will be readily accepted.

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 [5]. In order to make full use of the data, the campaigns customer conversion results and related independent parameters should be analysed. This process will further help the enterprise to make intelligent decisions for future campaigns and services.

II. RESEARCH QUESTIONS

The following research questions will be addressed in the proposal :

1) Compute the most important factors and predict the price of house from multiple parameters and establish a relationship.
2) Evaluate the decision of a customer to open a term-deposit after a bank marketing campaign.

III. INITIAL REVIEW

There are different data science methodologies/ standardized processes in the DM application such as: Knowledge Discovery in Databases (KDD), Sample - Explore - Modify - Model - Assess (SEMMA) and CRoss-Industry Standard Process for Data Mining (CRISP-DM). Steps involved in these processes are mostly similar with few additional or small changes [2]. For this research project KDD will be better data science methodology to incorporate the DM and ML concepts for the reason that we don't require complete domain/ business knowledge regarding banking and real-estate to create a model.

The house pricing of US market in Fairfax County, Virginia is calculated using various classification models, namely RIPPER, C4.5 Decision Tree, Naive Bayes and AdaBoost on particular kind of residential properties and town-houses predictors. Comparison of the accuracy is only performed on classification models [3]. Since in this model only classification is used, it has not incorporated some of the important quantitative continuous parameters, such as size, number of rooms and bathrooms, floors, age of house. Therefore, in the proposed research, regression modeling will be focused on to evaluate the price of houses.

To classify the screening candidates in an organization for a vacancy, classification models are used, namely KNN and Naive Bayes. Profile attributes of a candidate are various independent factors which are used to classify the candidate selection in the organization [6]. Bank marketing data set is of similar kind where dependent variable is of classification type with related parameters which can be distributed as factors,like job, marital status, month, and education. Classification type model will be a good fit to predict the decision of customer to open term-deposit based on other parameters. For the research, multiple classification models will be generated and the accuracy will be compared to get the best model.

IV. DATA SOURCES

Given below are the data sources used for the research questions for the proposal :

A. Sales of houses in King County, USA dataset

This dataset has real 21 613 records with 21 parameters about the price of house sale in King County (including Seattle), USA [7].

B. Computer generated house pricing dataset

This dataset is computer generated and contains 16 parameters of the house including price. It has about half a million records. [8].

C. Bank marketing campaigns dataset,Portugal bank

This dataset describes the status(yes/no) of clients opening a term-deposit in Portugal bank. It has 17 parameters and 45 211 rows [9].

V. MACHINE LEARNING METHODS

A. Multiple Linear Regression

Multiple Linear Regression (MLR) is the most common modeling technique for numeric data. It can be used to model linear variables, a polynomial combination, and interaction terms in the data. Additionally, forward, backward, mixed step approach can be used to predict the most suitable predictors for the dependent variable. It works on the principle, which assigns the coefficients with predictors in additive format such that the value of coefficient is equal to amount of change in the dependent variable with a unit change in independent variable [10].

B. Decision Trees

Decision Trees divide the data into subsets and then partitions into further subsets and so on till the data is sufficiently homogeneous. This modeling method is used to perform classification using C5.0, 1R and RIPPER algorithms, and regression using Model Trees [11].

C. k-Nearest Neighbor (k-NN)

The k-NN is one of the most simple and effective ML algorithms. It doesn't make any presumptions regarding the data distribution underlying and training phase is fast. It can be used both for regression and classification. Normalization of data before modelling is ideal to allow equal weight-age to all the predictors in the data. The concept behind k-NN is to locate k closest classified records to classify unlabelled records/ examples. Classification differs with different k values which are changed to predict the best model [11].

D. Naive Bayes

Naive Bayes method works on Bayes' Theorem, which works on the principal of probability of likelihood of an event based on previous results. This method is also fast, simple and very effective. Normal distribution of underlying data is assumed in this model. It also works well with missing and noisy data. It is only used for classification modelling [11].

E. Support Vector Machine

Support Vector Machine (SVM) plots data in multi-dimensional plane which represents records/ examples and parameter/ feature values. It can be treated as a surface which creates a boundary among data points to creates classification. It is mostly used to model the relationships which are highly complex. It combines the features of k-NN and Linear regression [11].

Data sets IV-A and IV-B contains multiple numeric/ quantitative predictors such as size, number of rooms and bathrooms, etc. Moreover, predicted/ dependent value is a continuous value and therefore, Multiple Linear Regression and Model Tress (Decision Tress regression) methodology will be a good fitting model for 1. Whereas, data set IV-C has most of the parameters of type factors, such as education, and martial status along with the yes/ no value classifying predicted values and thus classification machine learning methodology is better fit for 2.

VI. EVALUATION METHODS

A. Regression Model Evaluation

Below two statistics will be used to evaluate the regression model used for research question 1 and applied on data-sets IV-A and IV-B

1) R^2 and adjusted R^2 : Value of R^2 ranges from 0 to 1. Greater the value R^2, the better is the model. Adjusted R^2 is calculated on R^2 by penalizing it for every addition of predictor to the linear model.
2) Root Mean Square Error (RMSE) : Lower the value of RMSE, better is the fit of model.

B. Classification Model Evaluation

Below are the statistics that will be used to evaluate the performance of classification type model for research question 2 and applied on datasets IV-C. Values of following statistics ranges from 0 to 1. Greater the value, better the fit of the model.

1) Accuracy
2) Kappa
3) Sensitivity
4) Specificity
5) F-measure
6) Area under the curve (AUC)

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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.
[7] Harlfoxem. (2016) House sales in king county, usa. [Online]. Available: https://www.kaggle.com/harlfoxem/housesalesprediction
[8] A. Sleem. (2018) House pricing. [Online]. Available: https://www.kaggle.com/greenwing1985/housepricing
[9] 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
[10] J. Gareth, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 4th ed. Springer.
[11] L. Brett, *Machine Learning with R*, 2nd ed. Packt Publishing.