CAR PREDICTION POPULARITY

Kandula Rishitha, Sri.S.K.Alisha, Sri.V.Bhaskara Murthy
Mca Student, Senior Assistant Professor, Associate Professor
Dept Of Mca
B.V.Raju College, Bhimavaram

ABSTRACT-Today is a world of technology with a foreseen future of a machine reacting and thinking same as human. In this process of emerging Artificial Intelligence, Machine Learning, Knowledge Engineering, Deep Learning plays an essential role. In this paper, the problem is identified as regression or classification problem and here we have solved a real world problem of popularity prediction of a car company using machine learning approaches. Keywords—Machine Learning, Classification, Regression, Supervised Machine Learning, Logistic Regression, Random Forest , KNN.

I. INTRODUCTION

In the era which we live in, technology has a big impact on our lives. Artificial intelligence [6], knowledge engineering, Machine learning, Deep learning [4][5], Natural language processing[7][8] are emerging technologies which plays an important role in the leading projects of today's world. Artificial intelligence is an area or branch which aims or emphasizes on creating machine that works intelligently and their reactions is similar to that of human. In Artificial Intelligence, Machine learning is an essential and core part providing the ability of learning and improving by itself. The focus of this technique is on creation of programs which can pick the data and learn from it by itself. Earlier, statistician and developers worked together for predicting success, failure, future etc. of any product. This process led to delay of the product development and launch. Maintenance of such product in the changing technology and data is also one of the major challenges. Machine learning made this process easier and faster. There are various Machine learning algorithms broadly categorized into four paradigms:

• Supervised learning [7] [9] [10]: This learning algorithm provides a function so as to make predictions for output values, where process starts from analysis of a known training dataset. This algorithm can be applied to the past learned data to new data using labels so as to predict future events.

• Unsupervised learning: This algorithm is used on training dataset and informs which is neither classified nor labeled. It also studies to infer a function from a system to describe a hidden structure from unlabeled data. Clustering is an approach of unsupervised learning.

• Semi supervised learning [6] [11]: It takes the characteristics of both unsupervised learning and supervised learning. These algorithms uses small amount of labeled data and large amount of unlabeled data.

• Reinforcement [12]: In this algorithm, interaction is made to environment by actions and discovering errors. It allows machines and software agents in determining ideal behavior in a specific context such that performance could be maximized. Regression and Classification problems are types of problems in supervised learning. In classification, conclusion is drawn using values which are obtained by observation. A discrete output variable say y is approximated by this problem using a mapping function say f

on input variables say x. The output of classification is generally discrete but it can also be continuous for every class label in the form of probability. A regression problem has output variable as a real or continuous value. A continuous output variable say y is approximated by this problem using a mapping function say f on input variables say x. The output of regression is generally continuous but it can also be discrete for any class label in the form of an integer. A problem with many output variables is referred to multivariate regression problem. In this paper we will be focusing on a problem picked from hackerrank where a company is trying to launch a new car modified on the basis of the popular features of their existing cars. The popularity will be predicted using machine learning approach. It can be classified as regression problem especially a multivariate regression problem and the problem can be classified under supervised learning. Thus various supervised learning algorithms will be used for this prediction.

II. EXISTING SYSTEM

In paper "Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks

[1]", author has reviewed some classification algorithms such as random forest, gradient boosted trees, artificial neural network and logistic regression to predict 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) 463 stocks of the S&P 500. In order to study the predictability of these stocks, author has performed multiples of experiments with these classification algorithms. The obtained result of predicting future prices from the past available data was not up to the mark as the expected result, The author wanted to obtain. However, they successfully showed the vast growth in predictability of European and Asian indexes closed a little while back. In paper "Performance evaluation of predictive models for missing data imputation in weather data

[2]", author has suggested a new approach to manage the missing data in weather data by performing various tests with NCDC dataset to assess the prediction error of five methods: linear regression, SVM, random forest, KNN Implementation and kernel ridge. In order to handle the missing values of dataset they performed two actions: 1.removing the entire row which contains missing value and 2. Impute the missing data. They performed both the methods to handle the missing data and compared the observed result. In paper "Amazon EC2 Spot Price Prediction using Regression Random Forests

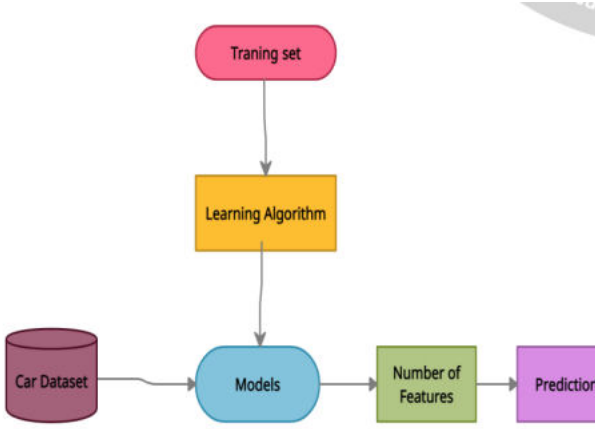
[3]", author has proposed Regression Random Forests (RRFs) model to forecast the Amazon EC2 Spot Price one week ahead and one month ahead. This prediction model would help in planning when to acquire the spot instance, the model also predicts the execution cost and it also suggests the user when to bid in order to minimize the execution cost

III. PROPOSED SYSTEM

The present system focuses on the introduction of some applicable AI-based strategies that can support existing standard methods of dealing with car popularity. Hence in the present work deep learning strategy is used. As a subset of machine learning. Dl consist of numerous layers of algorithms that provide a different interpretation of the data it feeds on. However, DL is mainly from ML because it Presents data in the system in a different manner. Whereas DL networks works by layers of Artificial Neural Network (ANN), ML algorithms are usually dependent on structured data. Unlike supervised learning which is that the task of learning a function mapping an input to an output on the premise of examples input-output pairs, unsupervised learning is marked by minimum

human supervision and will be described as a form of machine learning in search of undetected patterns in an exceedingly data set where no prior labels exist. DL can be extensively applied for car popularity; however, aims at finding the most effective possible solutions for car popularity related issues. With the aim of foregrounding the enhanced effectiveness of these strategies and techniques, their formation has been informed by. Therefore, this section presents ideas that can enhance and speed up ANN-based methods obtaining process to improve process methods.

III. SYSTEM ARCHITECTURE



IV. IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

MODULES DESCRIPTION:

• Buying Price

• Maintenance cost

• Number of doors

• Number of seats

• Luggage boot size

• Safety rating

Buying Price: The Buying price attribute is used to describe the buying price of the cars. It ranges from [1…4] where 1 represents the lowest price and 4 is representing highest price.

Maintenance Cost: The Maintenance Cost attribute is used to describe the maintenance cost of the cars. It ranges from [1…4] where 1 represents the lowest maintenance cost and 4 is representing highest maintenance cost.

Number of Doors: The number of Doors attribute is used to describe the number of doors in the car, and the values ranges from [2...5], where each value of number of doors represents the number of doors in the car.

Number of seats: The number of seats attribute is used to describe the number of seats in the car, and the values are [2, 4, 5], where each value of represents the number of seats in the car4. Luggage boot size: The luggage boot size attribute is used to denote the luggage boot size, and its values ranges from [1..3]. Value 1 smallest and 3 is largest luggage boot size.

Safety Rating: The Safety rating attribute is used to describe the safety rating of cars. Its value ranges from [1...3] where 1 represents low safety and 3 is high safety.

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not

a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. Three key considerations involved in the feasibility analysis are:

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

V. CONCLUSION

Machine Learning is a fast growing approach to solve real world problems. This paper focused on some of the supervised learning algorithms such as Logistic Regression, KNN, SVM and Random Forest for prediction popularity on a scaling measure of [1…4] for a car company. From table 1 it is clear that SVM is giving us the best result. Thus for future work, our focus would be on modifying SVM model used and will try to make the prediction more accurate. Also implementing the problem using deep learning deep learning and neural network algorithms will be our focus, as they provide more generalization of problems.

REFERENCES

[1] Jiao, Yang, and Jérémie Jakubowicz. "Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks." Big Data (Big Data), 2017 IEEE International Conference on. IEEE, 2017.

[2] Gad, Ibrahim, and B. R. Manjunatha. "Performance evaluation of predictive models for missing data imputation in weather data." Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on. IEEE, 2017.

[3] Khandelwal, Veena, Anand Chaturvedi, and Chandra Prakash Gupta. "Amazon EC2 Spot Price Prediction using Regression Random Forests." IEEE Transactions on Cloud Computing, 2017.

[4] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436..

[5] Le, Quoc V., Jiquan Ngiam, Adam Coates, Abhik Lahiri, Bobby Prochnow, and Andrew Y.

Ng. "On optimization methods for deep learning." In Proceedings of the 28th International Conference on International Conference on Machine Learning, pp. 265-272. Omnipress, 2011.

[6] Zhu, Xiaojin. "Semi-supervised learning literature survey." (2005).

[7] Olsson, Fredrik. "A literature survey of active machine learning in the context of natural language processing." (2009).

[8] Cambria, Erik, and White B. "Jumping NLP curves: A review of natural language processing research." IEEE Computational intelligence magazine 9.2 (2014): 48-57.

[9] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160 (2007): 3-24.

[10] Khan, A., Baharudin, B., Lee, L.H. and Khan, K., 2010. "A review of machine learning algorithms for text-documents classification." Journal of advances in information technology, 1(1), pp.4-20.

[11] Jiang J. "A literature survey on domain adaptation of statistical classifiers." URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey. 2008 Mar 6;3.

[12] Kaelbling, L.P., Littman, M.L. and Moore, A.W., 1996. "Reinforcement learning: A survey." Journal of artificial intelligence research, 4, pp.237-285

[13] Ban, Tao, Ruibin Zhang, Shaoning Pang, Abdolhossein Sarrafzadeh, and Daisuke Inoue. "Referential knn regression for financial time series forecasting." In International Conference on Neural Information Processing, pp. 601-608. Springer, Berlin, Heidelberg, 2013.

[14] Dutta, A., Bandopadhyay, G. and Sengupta, S., 2015. "Prediction of stock performance in indian stock market using logistic regression." International Journal of Business and Information, 7(1).

[15] Liaw, A. and Wiener, M. "Classification and regression by randomForest." R news (2002), 2(3), pp.18-22.

[16] Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. and Feuston, B.P. "Random forest: a classification and regression tool for compound classification and QSAR modeling." Journal of chemical information and computer sciences (2003), 43(6), pp.1947-1958.

[17] Smola, A.J. and Schölkopf, B. "A tutorial on support vector regression." Statistics and computing (2004), 14(3), pp.199-222.

[18] Gunn, S.R. "Support vector machines for classification and regression." ISIS technical report (1998), 14(1), pp.5-16.

[19] Williams, N., Zander, S. and Armitage, G. "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification." ACM SIGCOMM Computer Communication Review (2006), 36(5), pp.5-16.

[20] Willmott, C.J. and Matsuura, K. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance." Climate research (2005), 30(1), pp.79- 82