International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-07, Special Issue, MAY 2021

Car Popularity Prediction: A Machine Learning Approach

[1]Prof. Surekha Prakash Mali,[2]Mr.Dileep Santprasad Singh, [3]Mrs.Monika Rapte, [4]Mr.Jayesh Patil

[1]Asst.Professor,[2,3,4]UG Student,[1,2,3,4]Computer Engg. Dept. Shivajirao S. Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India.

[1]Surekhan46@gmail.com,[2]singhdileep834@gmail.com,[3]monikarapte93767@gmail.com, [4]jp8106424@gmail.com

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 today, 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 a crucial part in the leading papers 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 application 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 gives the information of the unclassified and unlabeled ones. Also it studies to infer a function from a system to elaborate an eclipsed structure belonging to the data which is classified as unlabeled. 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 minute number of data which is labeled and huge number of data which is unlabeled.
- Reinforcement [12]: In this algorithm, interaction is made to environment by actions and discovering errors. It permits the software agents plus the machines in determining ideal behavior precisely 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 assumed to be 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 this also can be continuous for every class label which lies in alignment with the format of probability. A regression problem has output variable as the real or otherwise a continuous value. A continuous output variable assumed to be 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 this can regarded as discrete for any class label as configuration 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. AIMS AND OBJECTIVE

a) Aim

The main aim of this paper to find the most popular cars.A car company has the data for all the cars that are present in the market. They are planning to introduce some new ones of their own, but first, they want to find out what would be the popularity of the new cars in the market based on each car's attributes.

Company will provide a dataset of cars along with the attributes of each car along with its popularity. Our task is to train a model that can predict the popularity of new cars based on the given attributes.

b) Objective

The main objective of this work is to find the most popular car in a market.an implementation based on Deep Learning that allows, to determine in a precise way the appearance of signs or anomalies capable of representing a case of car popularity. Providing a complete analysis of cars. Some specific objectives are:

- Increase the sales of cars in the market..
- Fill the void between AI-based methods and vehicals approaches and find popular cars.
- Fast finding a popular cars.

III. LITERATURE SURVEY

Paper 1: Predicting stock movement direction with machine learning

AUTHORS: Jiao, Yang, and JérémieJakubowicz

Forecasting the direction of stocks market has gained plenty of attention. Accurate forecasts Indeed has significant implications on strategies for trades. Surprisingly there is a very little data suggesting the importance of this topic been published. This research paper, briefly outlines the performance of the four classic classification algorithms: random forest, the gradient boosted trees, the artificial neural network and the logistic regression in prediction of 463 stocks belonging to the S&P 500. Thorough analysis of the predictability of such stocks were rectified by performing various experiments

The comparison between the 3 schemes namely the standard cross validation, the single validation and the sequential validation was done for validation of each of the prediction algorithm. As anticipated, the prediction of future prices of the stocks was not possible from their past, but we could definitely conclude that prediction of S&P 500 could be increased to an extend by considering the recent information of the recently closed European and Asian indexes.

Also towards the end we were able to interpreted that prediction of stocks from the financial sector was easiest amongst the other sectors.

Paper 2: Performance evaluation of predictive models for missing data imputation in weather data.

AUTHORS: Gad, Ibrahim, and B. R. Manjunatha

Real datasets can have missing values for a different reasons such as in data that were not kept on file and data corruption. Climate forecasting has a highly relevant effect in agricultural fields and industries sectors. The process of predicting climate conditions is required for different areas of life sectors. Handling missing data is significant becausea lot of machine learning algorithms performance are affected by missing values in addition, they do not support data with missing values. Various techniques have been used to process missing data problem and the most applied is removing any row that contains at least one missing value. Also, another approaches to solve missing data problems are to impute the missing data to yield a more complete dataset. In order to improve the accuracy of prediction with the climate data, missing value from dataset should be removed or imputed/predicted in the pre-processing phase before using the data for prediction or clustering in the analysis step. In this paper, we propose a new technique to handle missing values in weather data using machine learning algorithms by execute experiments with NCDC dataset to evaluate the prediction error of five methods namely the kernel ridge, linear regression, random forest, SVM imputation and KNN imputation procedure. The missing values were imputed using each method and compared to the observed value. Results of the proposed method were compared with existing techniques

Paper 3: On optimization methods for deep learning

AUTHORS:Le, Quoc V., JiquanNgiam, Adam Coates, AbhikLahiri, Bobby Prochnow, and Andrew Y. Ng

The transcendent approach in preparing profound learning advocates the utilization of stochastic angle plunge strategies . Notwithstanding its simplicity of execution, stochastic slope plunge are hard to tune and parallelize. These issues make it trying to create, troubleshoot and increase profound learning calculations with stochastic angle drop. In this paper, we show that more refined off-the-rack improvement techniques, for example, Limited memory BFGS and Conjugate angle with line search can essentially streamline and accelerate the way toward retraining profound calculations. In our investigations, the distinction between L-BFGS/CG and stochastic angle plummet are more articulated in the event that we think about algorithmic augmentations and equipment expansions . Our trials with dispersed streamlining support the utilization of L-BFGS with privately associated networks and convolutional neural organizations. Utilizing

L-BFGS, our convolutional network model accomplishes 0.69 % on the standard MNIST dataset. This is a condition of-theart result on MNIST among calculations that don't utilize mutilations or pretraining.

IV. EXISTINGSYSTEM

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.

V. COMPARTIVE STUDY

SR NO.	PAPER TITLE	AUTHOR / Publication NAME	METHOD	ADVANTAGE	DISADVANTAGE
1.	Car Price Prediction using Machine Learning Techniques.	EnisGegic, BecirIsakovic, Dino Keco, ZerinaMasetic, Jasmin Kevric	KNN, Random Forest	The final prediction model was integrated into Java application. Further more the model was evaluated using test data and the accuracy of 87.38% was obtained.	Different features like exteriors colors, door numbers, type of transmission, dimensions, safety, air condition, interior, whether it has navigation or not will also influence the car price.
2.	Predicting the Price of Used Cars using Machine Learning Techniques	Computer Science and Engineering Department, University of Mauritius, Reduit, MAURITIUS SameerchandPudaruth,	Random Forest · Ridge Regression · Lasso · K-Nearest Neighbor · XGBoost	Like multiple linear regression analysis, knearestneighbors, naïve bayes and decision trees have been used to make the predictions	A seemingly easy problem turned out to be indeed very difficult to resolve with high accuracy
3.	Car Sales Prediction Using Machine Learning Algorithms	Madhuvanthi.K, Nallakaruppan.M.K, Senthilkumar N C, Siva Rama Krishnan S,	methodology of analytic hierarchy process	Analytic hierarchy process, several machine learning algorithms such as linear regression, random forest we have inferred results that are so specific and accurate.	To get the best clusters use random tree and to get best accurate feature out of itwe process them in to random forest. Which is ultimately.
4	Vehicle Price Prediction using SVM Techniques	S.E.VISWAPRIYA, DURBAKA SAI SANDEEP SHARMA, GANDAVARAPU SATHYA KIRA	Random forest	The data must be collected using web scraper that was written in PHP programming language.	Large data sets are required,

Figure 1: Comparative Study

VI. PROBLEM STATEMENT

- System can handle only structured data and cannot handle unstructured data.
- Human intervention is required.
- Large data sets are required, which may not be available.
- Comparative less effective in feature detection.

VII. 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.

VIII. ALGORITHM

The general idea of working of Car popularity prediction algorithm is given as follow:

Algorithm Train Model

Step 1:

Load features

Train→pd.read_csv("train.csv")

Test→pd.read_csv("test.csv",names=columns)

Import SVM

X→train[columns]

y→train.popularity

clf→svm.SVC(kernel="rbf",C=300)

clf→clf.fit(X,y)

pred→clf.predict(test[columns])

Step 2:

Data Pre-processing: Pre-processing the data that is train.csv file.

Step 3:

Use the prediction algorithm select number of features of the car and predict the output.

Step 4:

Display the graph.

IX. MATHEMATICAL MODEL

A. KNN (K-Nearest Neighbor) [13]:

The other specialties of KNN is that it does not go through explicitly, which is other words means that the training phase is minimal and fast. Hence there is training data generalization as the all the data is generally needed in testing phases, in KNN, which is why KNN is referred as lazy algorithm Process of KNN-

Steps:

1. Output values to query scenario q of X nearest neighbors are stored in vector r = { r^1......r^x } by the following steps repeated X times in a loop.

a. From data set, next scenarios here I denotes ongoing iteration in the domain {1... P}.

b. If **t < d(q,s^t)**or t not set

Then **t ← d(q,s^t)** and f ←o^f

c. Do until all entries in data set are over.

d. Storing t in vector c and f in vector r.

2. Arithmetic mean is calculated for r-

$$\bar{r} = \frac{1}{X}\sum_{i=1}^{X} r_i$$

3. Return inverse r as output value for t.

B. Logistic Regression [14]:

For Predictive analysis, Logistic regression is a very satisfactory option. It can be used for dependent binary variable. The relationship between the independent (s) variable and dependent variable can be well explained and elaborated using this algorithm. It is one of the statistical method in which dependent binary containing data as

1. This algorithm aims at describing the relation among variables that are independent and discovers best fit model with binary characteristics of interest.

Logistics regression predicts a logit transformation by generating confidents of a formulation:

$$\text{logit(p)}= b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

here p denotes probability if characteristic of interest is present.

The logit transformation is defined as logged odds.

$$\text{Odds} = \frac{p}{(1-p)} = \frac{\textit{Probability of presence of characteristic}}{\textit{Probability of absence of characteristic}}$$

$$\text{And} \qquad \text{logit (p)} = \ln\left(\frac{p}{1-p}\right)$$

In logistic regression, estimation is made by choosing parameters that maximizes likelihood of observing the sample values.

C. Random Forest [15] [16]:

Random forest belongs to supervised classification algorithm which creates a forest and also makes it random somehow.

Larger the number of trees indicates more accurate is the results. The former is used for both classification and regression tasks. The classifier of random forest can handle missing values and it models for the categorical values.

Random Forest works in two stages:

1. First stage is creating a Random Forest.

a. Select K features randomly from m features.

b. Calculate node ‘d’, among ‘K’ features, using best split point.

c. Node is split into daughter nodes.

d. Do a to c until number of nodes reached is 1.

e. By repeating steps from a to d, n number of times to create n number of trees. Thus, a forest build.

2. Next stage is to make prediction using random forest classifier:

a. An outcome is predicted and stored by using testing features along with rules of each decision tree created randomly.

b. Votes are calculated for each predicted target.

2018 Fourth International Conference on Computing Communication Control and

Automation (ICCUBEA)

Final prediction is considered to be highest voted predicted target.

D. Support Vector Machine [17] [18]:

Support Vector Machine aka SVM is also a

Supervised machine learning algorithm primarily used for classification problems and regression problems.

It also aims at finding optimal separating hyper plane maximizing margin of the training data when the training data correctly classified. This form of algorithm has better generalization on the unseen data

X. SYSTEM ARCHITECTURE

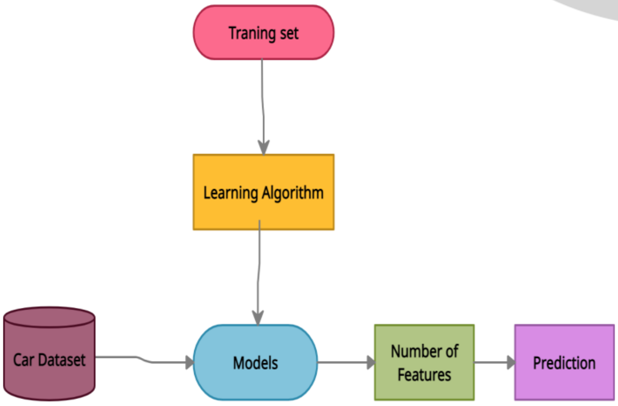

Fig.2: System Architecture

Description:

An architectural diagram is a diagram of a system which is utilized to abstract the overall outline of the software system and the relationships, boundaries and constraints, between its components. It helps to produce an overall view of the physical deployment of the software system and its evolution roadmap which makes it an essential part.

Learning Algorithm maps the input output pair. The learning algorithms analyze the training data and produce the interfered results. To develop our model we have used four algorithms namely, KNN, Logistic Regression, Random Forest and Support Vector Machine These steps further develop a model. Based on the number of features the model will predict the car popularity.

XI. ADVANATGES

1. Saves time and efforts
2. This system does effective prediction of car popularities.
3. Fast and accurate car popularities.

XII. DESIGN DETAILS

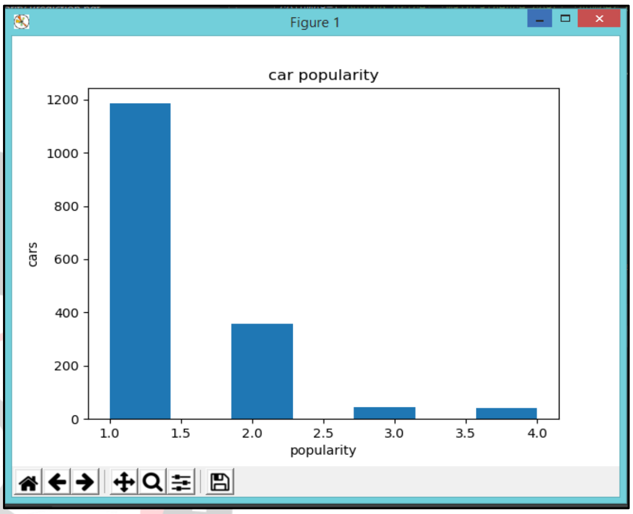

Figure 3: Bar graph of car popularitiy

Safety Rating of the cars is described by the Safety Rating. Here, the value ranges from 1 (low) safety to 3 (high) safety. Similarly, the popularity attribute decides the car popularity. The value ranges from 1 to 4, where 1 is for the unacceptable car, 2 is for an acceptable car, the value 3 represents a good car and lastly 4 represents the best car as show in figure 3 and figure 4.

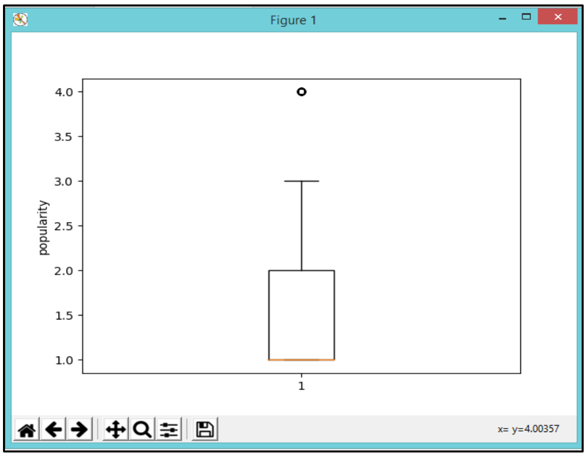

Figure 4: graph of car popularitiy

XIII. CONCLUSION

Thus, We have tried to implement the paper “Sunakshi Mamgain, Swati Vipsita , Srikant Kumar, KabitaManjari Nayak”, “Car popularity Prediction Using Machine Learning Approaches” IEEE 2020 and according to the implementation agriculture is the field which helps in economic growth of our country. But this is lacking behind in using new technologies of machine learning. Hence our farmers should know all the new technologies of machine learning and other new techniques. These techniques help in getting maximum yield of crops. Many techniques of machine learning are applied

on agriculture to improve yield rate of crops. These techniques also help in solving problems of agriculture. We can also get the accuracy of yield by checking for different methods. This paper helps in getting maximum yield rate of the crops. 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

REFERENCE

[1] Jiao, Yang, and JérémieJakubowicz. "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, YoshuaBengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436..

[5] Le, Quoc V., JiquanNgiam, Adam Coates, AbhikLahiri, Bobby Prochnow, and Andrew Y. Ng. "On optimization methods for deep learning."

[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 data 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, AbdolhosseinSarrafzadeh, and Daisuke Inoue. "Referential knn regression for monetary statistic prediction." In International Conference on Neural information science, pp. 601-608., 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 data, 7(1).

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