Car Popularity Prediction: A Machine Learning Approach

Sunakshi Mamgain
Department of Computer Science
M.Tech, I.I.I.T.-Bhubaneswar
Bhubaneeswar, India
a117009@iiit-bh.ac.in

Srikant Kumar
Department of Computer Science
M.Tech, I.I.I.T.-Bhubaneswar
Bhubaneeswar, India
a117008@iiit-bh.ac.in

Kabita Manjari Nayak
Department of Computer Science
M.Tech, I.I.I.T.-Bhubaneswar
Bhubaneeswar, India
a117003@iiit-bh.ac.in

Swati Vipsita
Department of Computer Science
I.I.I.T.-Bhubaneswar
Bhubaneeswar, India
swati@iiit-bh.ac.in

***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, Regression, Classification, Supervised Machine Learning, Logistic Regression, KNN, Random Forest.

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. RELATED WORKS

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

978-1-5386-5257-2/18/$31.00 ©2018 IEEE

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. ALGORITHMS

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

KNN has yet another specialty, it does not explicitly go through training phase or to say the training phase is minimal and fast. It also means that KNN does not use training data for generalization and all this data is generally needed in testing phase. Thus, KNN is often referred as lazy algorithm.

Process of KNN-

Assuming:

1. Data set is a matrix of dimension NXP.
2. P is scenarios s^1,….,s^P.
3. Each scenario contains N features; s^i={s_1^i,……,s_N^i}.
4. O is the vector of output values o^i for each scenario s^i, o={o^1…..o^P}.

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 s^i here I denotes ongoing iteration in the domain {1… P}.
   b. If $t < d(q, s^i)$ or t not set
      Then t ← $d(q, s^i)$ and f ← o^i.
   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-

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

3. Return $\bar{r}$ as output value for t.

B. *Logistic Regression [14]:*

Logistic regression is an appropriate predictive analysis. For a binary variable which is dependent, logistic regression is used. Using this algorithm a relation between independent variable(s) and dependent variable can be explained and data can also be described. It is statistical method in which variables which are dependent are binary containing data as 1. The aim of this algorithm is describing relation among variables which are independent and characteristics of interest which are binary by discovering a model which is best fitted. Logistic regression predicts a logit transformation by generating confidents of a formula:

$$\text{logit}(p) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k \qquad (2)$$

here p denotes probability if characteristic of interest is present.

The logit transformation is defined as logged odds.

$$\text{Odds} = \frac{p}{(1-p)} = \frac{\textit{Probability of presence of characteristic}}{\textit{Probability of absence of characteristic}} \qquad (3)$$

$$\text{And} \quad \text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \qquad (4)$$

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

C. *Random Forest [15] [16]:*

Random forest is a type of a supervised classification algorithm for creating a forest and making it random by some way.

Larger number of trees indicates more accuracy in results. Random forest is used for both classification and regression tasks. The classifier of random forest can handle missing values and can be modeled for 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 is build.
2. Next stage is to make prediction using random forest classifier:
   a. A 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.

Final prediction is considered to be highest voted predicted target.

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

Support Vector Machine also referred to as SVM is also a supervised machine learning algorithm used mostly for classification problems and also used for regression problems. SVM's objective is finding optimal separating hyper plane maximizing margin of the training data if it classifies training data correctly and this algorithm does generalization better on unseen data.

IV. EXPERIMENT DETAILS AND NUMERICAL SIMULATIONS

There are two data sets available in a .csv file which is comma separated file with useful information:

1. Train.csv -:
   This is a file that is used as training dataset whose each row provides information on each car. With values such as buying_price, maintenance_cost, number_of_doors, number_of_seats etc. Some of the attributes are explained as follows-
   a. 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.
   b. 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.
   c. 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.
   d. 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 car.
   e. 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.
   f. 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.
   g. popularity: The popularity attribute is used to describe the popularity of the cars. Its values ranges from [1…4] where 1 represents the unacceptable car, 2 represents an acceptable car, 3 represents a good car, and 4 represents the best car.

We have performed the experiment in python programming language. We have used pandas, numpy, matplotlib, seaborn, sklearn python libraries for solving the problem.
The snippet of training data is shown in fig 1.The schema of training data is shown in fig 2.

Brief description of training data is shown in fig 3.

	buying_price	maintainence_cost	number_of_doors	number_of_seats	luggage_boot_size	safety_rating	popularity
0	3	2	4	2	2	2	1
1	3	2	2	5	2	1	1
2	1	4	2	5	1	3	1
3	4	4	2	2	1	2	1
4	3	3	3	4	3	3	2

Fig 1: Training Data

```
In [5]: train.info()
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 1628 entries, 0 to 1627
        Data columns (total 7 columns):
        buying_price        1628 non-null int64
        maintainence_cost   1628 non-null int64
        number_of_doors     1628 non-null int64
        number_of_seats     1628 non-null int64
        luggage_boot_size   1628 non-null int64
        safety_rating       1628 non-null int64
        popularity          1628 non-null int64
        dtypes: int64(7)
        memory usage: 89.1 KB
```

Fig 2: Training Data Schema

```
In [6]: train.describe()
Out[6]:
```

	buying_price	maintainence_cost	number_of_doors	number_of_seats	luggage_boot_size	safety_rating	popularity
count	1628.000000	1628.000000	1628.000000	1628.000000	1628.000000	1628.000000	1628.000000
mean	2.532555	2.528256	3.493857	3.633292	1.987101	1.977887	1.348280
std	1.109626	1.116920	1.120557	1.257815	0.816520	0.819704	0.654766
min	1.000000	1.000000	2.000000	2.000000	1.000000	1.000000	1.000000
25%	2.000000	2.000000	2.000000	2.000000	1.000000	1.000000	1.000000
50%	3.000000	3.000000	3.000000	4.000000	2.000000	2.000000	1.000000
75%	4.000000	4.000000	4.250000	5.000000	3.000000	3.000000	2.000000
max	4.000000	4.000000	5.000000	5.000000	3.000000	3.000000	4.000000

Fig 3: Training Data Description

Training data visualization:

Fig 4 represents bar chart of parameter popularity where x axis represents popularity on the scale of 1 to 4 and y represents total count of cars belonging to a particular scaling parameter. Fig 5 represents hexplot of parameter popularity where x axis is representing safety_rating on the scale of 1 to 3 and y axis is representing popularity on the scale of 1 to 4. Fig 6 represents stacked plot of parameter popularity where x axis is representing buying_price, maintainence_cost on the scale of 1 to 4 and y axis is representing safety_rating, popularity on the scale of 0 to 3.5.

2. Test.csv -:

It is the test dataset of cars along with above attributes excluding popularity. The goal is to predict the popularity of cars of test dataset based on their remaining attributes.

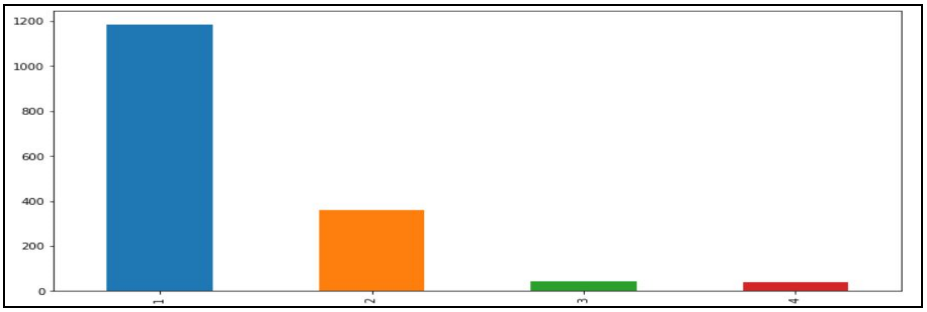

Fig 4: Bar Chart representation of Popularity parameter

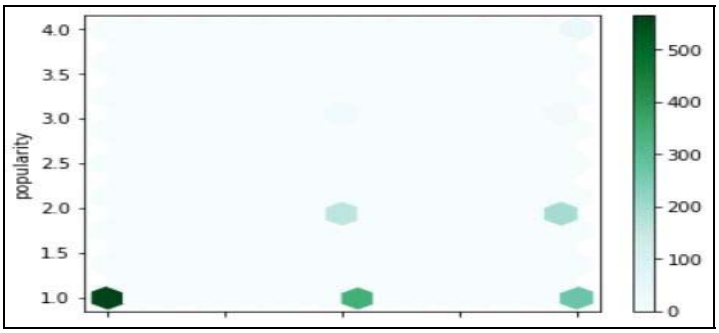

Fig 5: Hexplot of popularity

Fig 6: Stacked Plot of parameter popularity

V. RESULT AND DISCUSSION

After executing the Machine Learning Algorithm the next step is to find out the effectiveness of model based on various performance metrics. Different performance metrics are used for different Machine Learning Algorithms. For example: For classification we use different performance metrics [19] such as Accuracy, Cross Validation, Precision, Recall, and f1 Score. If the machine learning algorithm is used for prediction (for example: stock price prediction, housing prediction and like in our case car popularity prediction) we use Root Mean Square Error (RMSE) [20], Mean Square Error (MSE) [20]. Because of absence of output data, we are unable to measure the performance of the Machine Learning Algorithms we applied in this problem. However, we have stored the predicted output values in .csv file we received after performing the algorithms we implemented in this paper. We have calculated the accuracy of the machine learning models we implemented which is shown in table 1.

VI. CONCLUSION AND FUTURE WORK

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 JC, Sheridan RP, and Feuston BP "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.*

TABLE I. TRAING TESTING ACCURACY OF MODELS

Model	Training Accuracy	Test Accuracy
KNN	0.97	0.94
Logistic Regression	0.83	0.99
Random Forest	**0.86**	**0.98**
SVM	0.97	0.99