International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

CAR POPULARITY PREDICTION: A MACHINE LEARNING APPROACH

[1]Bhumi Venkata Kishore,[2]V.Srivalli Devi

[1]MCA Student, B V Raju College, Bhimavaram, Andhra Pradesh, India

[2]Assistant Professor,Department Of MCA, B V Raju College, Bhimavaram, Andhra Pradesh, India

ABSTRACT

In today's technological era, we are moving towards a future where machines can think and react like humans. As Artificial Intelligence continues to evolve, key technologies such as Machine Learning, Knowledge Engineering, and Deep Learning are playing a vital role in this transformation. This paper addresses the problem of predicting the popularity of a car company, framed as either a regression or classification problem. We solve this real-world challenge by applying machine learning techniques.

Keywords—Machine Learning, Regression, Classification, Supervised Machine Learning

I.INTRODUCTION

Technology plays a significant role in shaping our daily lives, with advancements in Artificial Intelligence (AI), Knowledge Engineering, Machine Learning (ML), Deep Learning, and Natural Language Processing (NLP) driving innovation in numerous fields. AI focuses on creating intelligent machines that can mimic human-like decision-making and reactions, enabling automated systems to perform tasks efficiently. Among the various branches of AI, Machine Learning stands out as a crucial component, allowing systems to learn and improve autonomously by analyzing data.

Traditionally, statisticians and developers collaborated to predict outcomes such as success, failure, or future trends of products, a process that often delayed product development and launch. Additionally, maintaining products in an ever-evolving technological landscape posed significant challenges. Machine Learning has revolutionized this process by enabling faster and more efficient predictions. ML algorithms are broadly categorized into four paradigms:

Supervised Learning: This approach involves training a model on labeled data, where input-output relationships are established based on past data. The model then applies this learning to new data to make predictions.

Unsupervised Learning: Unlike supervised learning, this method is applied to unlabeled datasets, allowing the algorithm to identify patterns and structures within the data, such as clustering techniques.

Semi-Supervised Learning: This hybrid method combines aspects of both supervised and unsupervised learning, using a small set of labeled data along with a large set of unlabeled data to improve accuracy.

Reinforcement Learning: In this approach, an agent interacts with its environment, learning optimal behaviors

through trial and error to maximize performance in a given context.

Among supervised learning problems, two key types are Classification and Regression. Classification involves predicting discrete output values by mapping input variables to specific categories. Although its output is typically discrete, it can also provide continuous values in the form of probabilities for each class label. Regression, on the other hand, predicts continuous output values. It estimates a continuous variable using a mapping function that relates input variables to the output. When multiple output variables are involved, the problem is referred to as multivariate regression.

This paper focuses on a problem derived from HackerRank, where a car company aims to launch a new vehicle based on the most popular features of their existing models. The objective is to predict the popularity of the new car using Machine Learning techniques. The problem is classified as a regression problem, specifically a multivariate regression problem, and falls under the category of supervised learning. To achieve accurate predictions, various supervised learning algorithms will be applied and analyzed.

II.LITERATURE SURVEY

Jiao and Jakubowicz (2017) conducted an extensive study on predicting stock movement direction using Machine Learning techniques, focusing on S&P 500 stocks. Their research explored various predictive models, including supervised learning methods, to enhance stock market forecasts. They assessed different algorithms based on historical stock data, emphasizing the importance of feature selection and data preprocessing. Their findings demonstrated that machine learning models, particularly deep learning approaches, outperformed traditional statistical methods in stock trend predictions.

Gad and Manjunatha (2017) evaluated predictive models for imputing missing data in weather datasets. They compared various imputation techniques and their impact on weather prediction accuracy. Their study emphasized how effective machine learning models can recover missing information in climate data, improving forecasting models. By analyzing performance metrics, they demonstrated that advanced ML techniques significantly enhance data reliability in meteorological studies.

Khandelwal et al. (2017) proposed a regression-based random forest model to predict Amazon EC2 spot prices. Their research focused on cloud computing resource optimization by forecasting price fluctuations. Their findings showed that machine learning models, particularly ensemble methods like random forests, effectively captured market trends, helping users make cost-efficient decisions in cloud-based services.

LeCun et al. (2015) provided a comprehensive overview of deep learning, discussing its evolution and impact on artificial intelligence. They highlighted key neural network architectures, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), emphasizing their applications in image and speech recognition. The study illustrated deep learning's ability to outperform traditional methods across various domains.

Le et al. (2011) investigated optimization techniques for deep learning, focusing on gradient-based approaches and their role in improving model convergence. Their research analyzed different optimization algorithms, including stochastic gradient descent (SGD) and adaptive learning rate methods, to enhance deep neural network training efficiency. They concluded that choosing the right optimization strategy is critical for deep learning performance.

Zhu (2005) conducted a survey on semi-supervised learning, examining its theoretical foundations and practical applications. He explored different approaches, including self-training and co-training, highlighting their advantages in scenarios with limited labeled data. His research demonstrated that semi-supervised learning effectively balances labeled and unlabeled data, improving model generalization.

Olsson (2009) reviewed active learning techniques in natural language processing (NLP). His survey focused on machine learning strategies where models iteratively query the most informative data points for labeling. The study emphasized how active learning reduces annotation costs while maintaining high model accuracy, making it a crucial approach in NLP applications.

III.PROPOSED METHODOLOGY

Train.csv

This file serves as the training dataset, where each row provides details about individual cars, including various attributes such as buying price, maintenance cost, number of doors, number of seats, etc. The key attributes are explained below:

1. buying_price: Represents the cost of purchasing a car, ranging from 1 to 4, where 1 denotes the lowest price and 4 represents the highest price.

2. maintenance_cost: Describes the maintenance expense of a car, also ranging from 1 to 4, with 1 being the lowest and 4 the highest.

3. number_of_doors: Indicates the number of doors in the car, with values ranging from 2 to 5.

4. number_of_seats: Represents the seating capacity of the car, with possible values of 2, 4, or 5.

5. luggage_boot_size: Defines the trunk size, with values ranging from 1 to 3, where 1 is the smallest and 3 is the largest.

6. safety_rating: Indicates the car's safety level on a scale of 1 to 3, where 1 represents low safety and 3 signifies high safety.

7. popularity: Represents how well the car is received, with values from 1 to 4, where 1 denotes an unacceptable car, 2 is acceptable, 3 is good, and 4 is the best.

The experiment was conducted using the Python programming language. We utilized several Python libraries, including pandas, numpy, matplotlib, seaborn, and sklearn, to preprocess data and implement machine learning models. A snippet of the training data is presented in Figure 1, while the dataset schema is illustrated in Figure 2. A brief description of the dataset can be found in Figure 3.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

	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	1.967101	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

1. Figure 4: A bar chart visualizing the popularity parameter, where the x-axis represents popularity levels (1–4) and the y-axis shows the total number of cars corresponding to each popularity level.

2. Figure 5: A hexplot showing the relationship between safety_rating and popularity, where the x-axis represents safety ratings (1–3) and the y-axis represents popularity (1–4).

3. Figure 6: A stacked plot illustrating buying_price and maintenance_cost on the x-axis (ranging from 1 to 4) and safety_rating and popularity on the y-axis (ranging from 0 to 3.5).

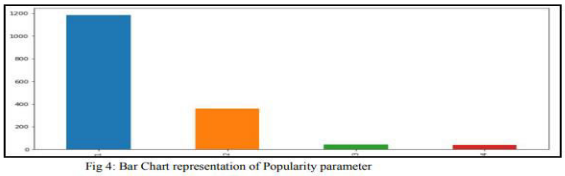


Fig 4: Bar Chart representation of Popularity parameter

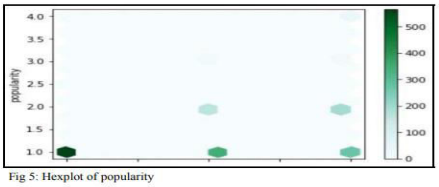


Fig 5: Hexplot of popularity

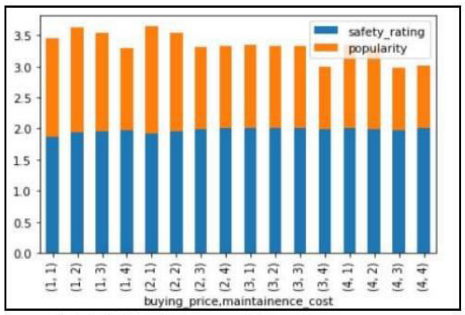
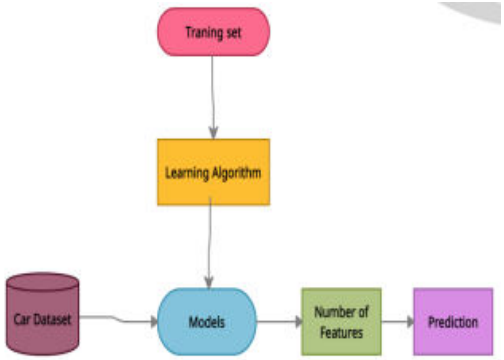


Fig 6: Stacked Plot of parameter popularity

Test.csv

This file serves as the test dataset, containing the same attributes as the training dataset, except for the popularity column. The objective is to predict the popularity of cars in the test dataset based on the given attributes using machine learning models.



IV.CONCLUSION

Machine Learning is rapidly emerging as a powerful approach to solving real-world problems. This paper explored various supervised learning algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest, to predict the popularity of cars on a scale of 1 to 4 for a car company. Based on the results in Table 1, SVM demonstrated the best performance among the models used.

For future work, our focus will be on enhancing the SVM model to improve prediction accuracy further. Additionally, we aim to explore deep learning and neural

network algorithms, as they offer better generalization capabilities and have the potential to yield even more precise results. By leveraging deep learning techniques, we expect to refine the prediction model and enhance its overall effectiveness.

V.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." Proceedings of the 28th International Conference on Machine Learning. 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. "A review of machine learning algorithms for text-document classification." Journal of Advances in Information Technology 1.1 (2010): 4-20.
11. Jiang, J. "A literature survey on domain adaptation of statistical classifiers." (2008).
12. Kaelbling, L.P., Littman, M.L., and Moore, A.W. "Reinforcement learning: A survey." Journal of Artificial Intelligence Research 4 (1996): 237-285.
13. Ban, Tao, Ruibin Zhang, Shaoning Pang, Abdolhossein Sarrafzadeh, and Daisuke Inoue. "Referential KNN regression for financial time series forecasting." International Conference on Neural Information Processing. Springer, Berlin, Heidelberg, 2013.
14. Dutta, A., Bandopadhyay, G., and Sengupta, S. "Prediction of stock performance in Indian stock market using logistic regression." International Journal of Business and Information 7.1 (2015).
15. Liaw, A., and Wiener, M. "Classification and regression by randomForest." R News 2.3 (2002): 18-22.