# Car Popularity Prediction : A Machine Learning Approach

K V V S Trinadh Naidu[1], T.Sushma Reddy[2], T.B.N.L.Keerthana[3], P.S.L.S.Mounika[4], K.Swarna Bharathi[5]
*[1,2,3,4,5]Pragati Engineering College*

***Abstract* - Car Popularity Prediction: A machine Learning Approach 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. The focus of this technique is on creation of programs which can pick the data and learn from it by itself. 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. This project contains what and how the car predicting model works with the help of machine learning and which dataset is used in our proposed model.**

## I.INTRODUCTION

In the era which we live in, technology has a big impact on our lives. Artificial intelligence, knowledge engineering, Machine learning, Deep learning, Natural language processing 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 : 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: 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 : 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 LITERATURE SURVEY

Stocks movement direction forecasting has received a lot of attention. Indeed, being able to make accurate forecasts has strong implications on trading strategies. Surprisingly enough little has been published, relatively to the importance of the topic. In this paper, we reviewed how well four classic classification algorithms: random forest, gradient boosted trees, artificial neural network and logistic regression perform in predicting 463 stocks of the S&P 500. Several experiments were conduced to thoroughly study the predictability of these stocks. To validate each prediction algorithm, three schemes we compared: standard cross validation, sequential validation and single validation. As expected, we were not able to predict stocks future prices from their past. However, unexpectedly, we were able to show that taking into account recent information - such as recently closed European and Asian indexes - to predict S&P 500 can lead to a vast increase in predictability. Moreover, we also found out that, among various sectors, financial sector stocks are comparatively more easy to predict than others.

## III EXISTING SYSTEM

Data was collected from <<petites announces>> found in daily newspapers such as L'Express and Le Defi . We made sure that all the data was collecte in less than one month interval as time itself could have an appreciable impact on the price of cars. In Mauritius, seasonal patterns is not really a problem as this does not really affect the purchase or selling of cars. The following data was collected for each car: make, model, volume of cylinder (funnily this is usually considered same as horsepower in Mauritius), mileage in km, year of manufacture, paint colour, manual/automatic and price. Only cars which had their price listed were recorded.Because many of the columns were sparse they were removed. Thus, paint colour and manual/automatic features were removed. The data was then further tweaked to remove records in which either the age (year) or the cylinder volume was not available. Model was also removed as it would have been extremely difficult to get enough records for all the variety of car models that exist. Although data for mileage was sparse, it was kept as it is considered to be a key factor in determining the price of used cars,

## IV PROPOSED SYSTEM

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. A problem with many output variables is referred to multivariate regression problem.

## V REQUIREMENT ANALYSIS

The project involved analyzing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

### INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy.

Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for

immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

- Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
- Select methods for presenting information.
- Create document, report, or other formats that contain information produced by the system

## VI IMPLEMENTATION

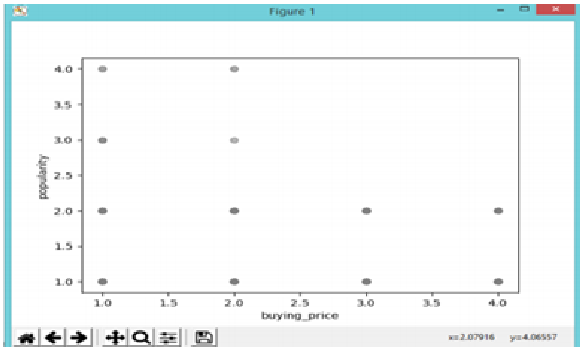
Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation 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:

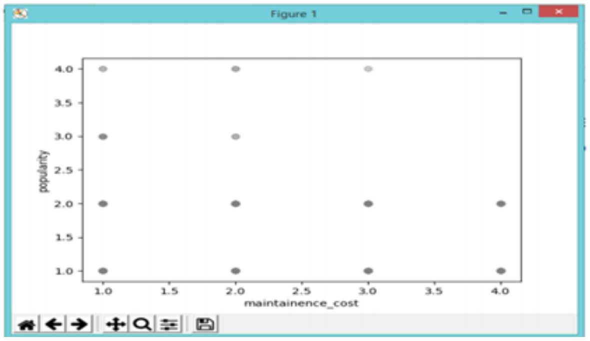
The Buying price attribute is used to describe the buying price of the cars. It ranges from [1…4] where 1 represents the lowest price and 4 is representing highest price.



Maintenance Cost:

The Maintenance Cost attribute is used to describe the maintenance cost of the cars. It ranges

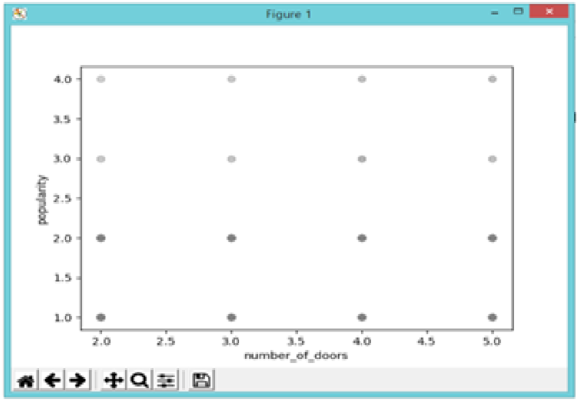
from [1…4] where 1 represents the lowest maintenance cost and 4 is representing highest maintenance cost.



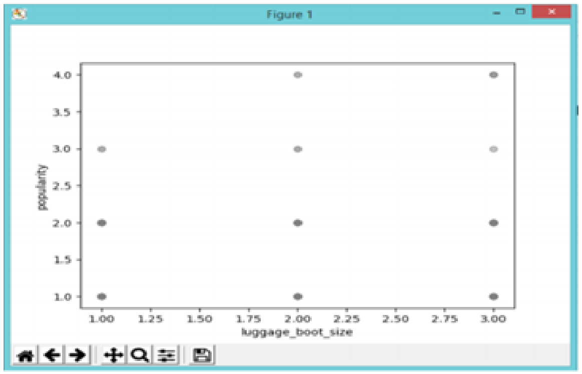
Number of Doors:

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.

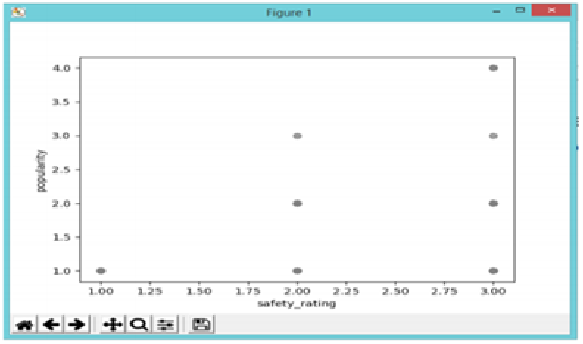
Number of seats:
The number of seats attribute is used to describe the number of seats in the car, and the values are [2, 4, 5], where each value of represents the number of seats in the car4.



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.



## VII SYSTEM STUDY

### 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.

## VIII. CONCLUSION

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. Hence we can improve the performance by checking the accuracy between different crops. Sensor technologies are implemented in many farming sectors. This paper helps in getting maximum yield rate of the crops. Also helps in selecting proper crop for their selected land and selected season. These techniques will solve the problems of farmers in agriculture field. This will help in improving the economic growth of our country. 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.

## IX ACKNOWLEDGMENT

I would like to express my gratitude to all the people behind the screen who helped me to transform an idea into a real application.

I would like to thank my internal guide Mr. K V V S Trinadh Naidu for his technical guidance, constant encouragement and support in carrying out my project. The satisfaction and euphoria that accompany the successful completion of the task would be great but incomplete without the mention of the people who made it possible with their constant guidance and encouragement crowns all the efforts with success.

## 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."

[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.