International Journal of Gender, Science and Technology
ISSN: 2040-0748

UGC Care Group I Journal
Vol-8 Issue-02 Aug 2019

Machine Learning for the Prediction of Car Popularity

N.Neelima Priyanka [1], Amritha Mishra [2]
Assoicate Professor [1], Assistant professor [2]
Department of IT, SRK INSTITUTE OF TECHNOLOGY ENIKEPADU VIJAYAWADA
Mail Id : priyankanutulapati@gmail.com, Mail id : amrithapandey@gmail.com

Abstract

Human-like computers are becoming a reality in today's technologically sophisticated world. Deep Learning is one of the most essential components of the constantly expanding area of artificial intelligence. According to this study, the popularity of a car manufacturer may be predicted using machine learning methods, which classifies the problem as either a regression or a classification issue.

INTRODUCTION

In today's world, technology has a profound effect on our daily routines. Artificial Intelligence (AI), Knowledge Engineering (KE), Machine Learning (ML), and Deep Learning (DL) are some of the most essential technologies in the world today (DL). Natural language processing[7][8] is another method that is gaining popularity. AI researchers are working hard to create robots that can think and behave just like us. Machine learning, which allows computers to learn and evolve on their own, is a key component of AI. Programs that can learn from their own data and use that knowledge in new settings are the emphasis of this method. There was a time when statisticians and engineers worked together to predict a product's success or failure. As a result of this approach, the product's conception and introduction were both delayed. It becomes more difficult to maintain a product like this as data and technology evolve. Machine learning made this process more efficient and less time consuming. There are four major groups of machine learning algorithms: In the case of supervised learning, it may be used to predict output values from a known training dataset.

A real or continuous output variable may be used in a regression issue. In this issue, a mapping function f on input variables x is utilised to approximate a continuous output variable, say y. Each class with an integer as a label may have its regression output be continuous or discrete. Multivariate regression is used when there are a number of possible outcomes to consider. We'll take a look at how a car company uses a hackerrank challenge to advertise a new model.

based on the most popular models in their current lineup. A product's popularity may be predicted with some accuracy thanks to machine learning. It may be considered an example of supervised learning in the context of regression and learning. This forecast will employ a variety of supervised learning approaches.

EXISTING SYSTEM:

An extensive study of 463 S&P 500 stocks was conducted for the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) to examine various classification algorithms, including random forest and gradient boosted trees, artificial neural networks, and logistic regression, in order to forecast their movement. Using these categorization approaches, the author has carried out a number of studies in order to better understand how dependable these stocks really are. When the author attempted to estimate future pricing based on historical data, his expectations were not satisfied. However, the recently closed European and Asian indices were shown to have a significant increase in predictability.

Authors used the NCDC dataset to assess the accuracy of five methods: linear regression, SVM, random forest, KNN Implementation, and Kernel Ridge in their article titled "Performance Evaluation of Predictive Models for Missing Data Imputation in Meteorological Data." After erasing the whole row containing the omitted value, the dataset had its missing data re-imputed. Results from both strategies were compared to determine which was more effective. For the next week and month, Amazon EC2 Spot Price may be predicted with the use of a model that employs regression random forests, as described in the paper "Amazon EC2 Spot Price Prediction Using RRFs" (RRFs). In addition to predicting the execution cost, our model advises users on when to bid to minimise the cost of executing a given task.

LITERATURE REVIEW

"Predicting Stock Movement Direction Using Machine Learning: An Extensive Study of the S&P 500," JAKUBOWICZ, JIAO, and JERÉMIE, IEEE INTERNATIONAL CKS 2017,

"Big Data" is a phrase used in the IT industry to describe this phenomenon. Stock market forecasting has been more popular in recent years. Even yet, being able to effectively forecast future market circumstances is an important aspect of trading strategy. Given its importance, this topic has gotten much less attention than it should have. 463 S&P 500 firms were predicted using random forests, gradient-boosted trees, artificial neural networks, and logistic regression. These stocks were put through a series of tests to see whether they could predict the future. Prediction strategies were validated using cross-validation, sequential validation, and single validation approaches. Equities in the financial industry are more predictable than those in other industries, according to our research. [1] There is no doubt that the stock market is dynamic and non-stationary [2]. If you think about it, investing choices are affected by a variety of variables that are not restricted to politics and company policy but also the economy and commodities prices. Assuming the Efficient Market Hypothesis [4] is correct and asset prices are fair and promptly adjust to reflect all previous and present information, stock price fluctuations should follow a random walk pattern.. If this theory is right, it will be difficult to predict the market. Several markets have put the EMH hypothesis to the test. However, at times the results seem to be discordant. The random walk model has been around since the 1960s.

Forecasting models with missing meteorological data input were investigated at the ICACCI 2017 International Conference on Advancements in Computing, Communications, and Informatics (ICACCI).

Many reasons can lead to missing values in real-world datasets, including data that has been deleted or changed by accident.. Agriculture and business both depend on accurate weather forecasts. Forecasters of the weather, for example, must be able to reliably anticipate the weather. Several machine learning methods fail when a dataset includes missing values. By eliminating all rows with at least one missing value, missing data may be rectified. Imputed data may be used in addition to actual data while filling in the gaps. It may be able to remove or predict missing data points as part of the pre-processing. After that, you may utilise the data to generate predictions or organise it into categories. Weather data is analysed using the National Climatic Data Center (NCDC) data as well as the Kernel Ridge approach and several statistical tools. Each method's efficacy was assessed by comparing the imputed values to the true ones. Contrasts between the new ideas and the current practises were evident throughout the meeting. Missing data is a common problem in real datasets and statistical research. It is impossible to generalise about the quantity of data that is missing across datasets. The quantity of data that is blank in a dataset varies substantially. There is a 1-5 percent tolerance for missing data, therefore "trivial" data is acceptable. Data mining and statistical analysis are significantly impacted by rates larger than 15% [4, 12]. Filling in the blanks is standard procedure when there are no missing values in a dataset. Estimating the value of missing data is possible via imputation. [11] Nonparametric or parametric regression may be used for imputation. It's important to note that data imputation and data analysis are two separate activities. Missing values are often re-implemented from partial rows of data in most imputation methods. You may find the closest neighbours of a missing value by using KNN imputation with the available data [8, 24]. Data recovery through case removal is the most usual method. All entries having a 0 value have now been removed from the database. Parameter estimation may be done using maximum likelihood methods. In many cases, parameter estimation techniques outperform case elimination methods because they can utilise the entire dataset.

CHANDRA PRAKASH GUPTA, ANAND CHATURVEDI, AND VEENA KHANDELWAL. In IEEE Transactions on Cloud Computing, "Amazon EC2 SPOT PRICE PREDICTION USING RANDOM FORESTS" is published.

To auction off extra capacity, Amazon EC2 announced "Spot Instances" in December 2009. Despite its low costs, the cloud spot market is underutilised. Due to the fluctuating nature of the spot price, out-of-bid failure is common. The complexity of bidding makes buyers wary of using spot timings. Prices for one week and one day are predicted using a regression random forest (RRFs) model. Cloud clients may prevent out-of-bid failure by predicting when to buy spot instances and evaluating execution costs. We utilised Amazon EC2 spot pricing data from the previous year to properly forecast future expenditures. Non-parametric machine learning models have been demonstrated to be less accurate than RRF spot price projections in studies. MAPE, MCPE, OBError, and speed may be used to quantify prediction accuracy. Statistics reveal that 66% to 92 percent of predictions are produced with MAPE = 10% and 35% to 81 percent with MCPE = 15% when forecasting one day in the future. MAPE is less than 15% in 71% to 96% of one-week projections. The best course of action is to

overestimate the number of resources you'll need to satisfy peak customer demand if you have no idea how many there are. Large portions of cloud resources sat unused at non-peak times because of an overestimation of demand. When cloud service providers allocate resources, they take into account the capacity of data centres as well as the needs of individual customers. It is easier to allocate resources because there is a wide range of users with different needs. Today's usage-based pricing model has resulted in a shift in demand for cloud resources. As a result, the use of strategies that allow for a wide range of price options is critical. EC2 announced spot pricing in December 2009 to reduce operating costs, combat underutilization of its resources, and increase profit margins. Spot instances and on-demand instances both offer a broad variety of CPU, memory, storage, and networking capabilities for virtual machines. In the market for spot instances (AWS), Amazon Web Services (AWS) is a new contender (AWS). On September 8, 2015, Google Compute Engine enabled preemptible virtual machines for workloads that may be delayed and are also fault tolerant. Google Preemptible VMs have a predetermined price for spot instances (SIs), and the lowest bidder wins. The dynamic pricing of Elastic Compute Cloud (EC2) spot instances sets it apart from the competition. Customers should expect low-cost utility computing using spot instances, notwithstanding the fact that Amazon EC2 can go down at any moment. When evaluating the reliability of a spot instance, look at its market price and the highest bid submitted by a single user (limited by their hourly budget). Prices for spot instances vary as demand (bids from users) and supply (data centres) alter in real time (resource availability). A user may, for example, decide on the reliability and cost of running a spot instance bid. On-demand instances might cost up to eight times as much as regular instances. The on-demand charges of cloud data centres, on the other hand, are seldom higher than the spot rates. A lower spot price is achieved when there are fewer bids competing for the same time slot.

PROBLEM STATEMENT:

- In order to process unstructured data, the system requires human interaction.
- Some of the data we need may not be accessible at this time.
- To detect characteristics, it is less effective.

PROPOSED SYSTEM:

Traditional approaches are being complemented with AI as autos become more common. As a result, deep learning is used in this project. This is a field of artificial intelligence that is included in the definition of machine learning. Each new piece of data that Dl gets adds another layer of algorithmic reasoning to the system. DL, on the other hand, largely relies on machine learning when it comes to data visualisation (ML). Structured data is the emphasis of machine learning (ML) rather than deep-learning (DL), which incorporates ANN layers. An output function is being trained in the context of supervised learning when an input is fed into the function. In contrast, unsupervised learning is good for discovering patterns in large datasets. DL may be used in a variety of ways to boost a car's popularity, but its major focus is on finding the most efficient options. They were created to draw attention to the improved efficacy of certain techniques and procedures. That is why this part focuses on principles that speed up and improve the acquisition of an ANN-based process optimization method.

REQUIREMENT ANALYSIS

In order to make it more user-friendly, the team studied the designs of rival programmes. There has to be a quick transition from one screen to the next and a minimum amount of text to accomplish this aim. A suitable browser version must be selected since not all browsers support the same version.

OBJECTIVES

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3.When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

INPUT DESIGN

The user interface of an information system is known as a "browser." By reading from a document, or by having people directly enter the data, computers can be used to store information. A portion of this is defining the data preparation needs and processes Keeping input under control, eliminating mistakes, and streamlining the process are all essential for making the process as simple as possible. The input technique protects users' privacy while also providing security and ease of use. The following factors were taken into consideration:

This information serves no use.

- The most effective way to organise this data is through coding or categorising it.
- Operators may provide each other feedback on their performance by conversing with each other.
- Input validation and error handling procedures.

HARDWARE REQUIREMENTS:

System	:	Intel i3
Hard Disk	:	1 TB.
Monitor	:	14' Colour Monitor.
Mouse	:	Optical Mouse.
Ram	:	4GB.

SOFTWARE REQUIREMENTS:

Operating system	:	Windows 10.
Coding Language	:	Python.
Front-End	:	Html. CSS
Designing	:	Html,css,javascript.
Data Base	:	SQLite.

CONCLUSION AND FUTURE WORK

In the actual world, machine learning is becoming increasingly difficult to implement. This study tested the accuracy with which supervised learning algorithms including Logistic Regression, KNN, SVM, and Random Forest can predict a car company's popularity using a randomly created scale of [1...4]. No question about it, SVM is the finest. We want to enhance the SVM model in the future so that we can make more accurate predictions about the future. Deep learning and neural networks, on the other hand, offer a far broader range of applications.

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