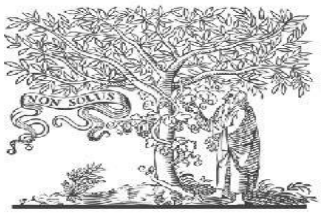
International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT



ELSEVIER
SSRN

IJIEMR Transactions, online available on 12th Jan 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue=ISSUE-01

DOI: 10.48047/IJIEMR/V11/I01/04

Title **COMPARATIVE ANALYSIS OF LINEAR REGRESSION, RANDOM FOREST AND SUPPORT VECTOR MACHINE USING DATASET**

Volume 11, Issue 01, Pages: 30-35

Paper Authors
Dr. Jitendra Singh Kushwah, **Dr.Rishi Soni**, **Dr.Aditya Vidyarthi**, **Shivanshu Ojha**

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COMPARATIVE ANALYSIS OF LINEAR REGRESSION, RANDOM FOREST AND SUPPORT VECTOR MACHINE USING DATASET

Dr. Jitendra Singh Kushwah, PhD
Department of Computer Science & Engineering
Institute of Technology and Management, Gwalior, Madhya Pradesh, India
Dr.Rishi Soni, PhD
Department of Computer Science & Engineering
Institute of Technology and Management, Gwalior, Madhya Pradesh, India
Dr.Aditya Vidyarthi, PhD
Department of Information Technology
Institute of Technology and Management, Gwalior, Madhya Pradesh, India
Shivanshu Ojha, B.Tech Student
Department of Computer Science & Engineering
Institute of Technology and Management, Gwalior, Madhya Pradesh, India

ABSTRACT:
Machine Learning computations can sort out how to perform basic tasks by summarizing from delineations. This exploration targets looking at changed calculations utilized in Machine Learning. Machine Learning can be both experience and clarification-based learning. In this concentrate, most well-known calculations were utilized like Linear Regression (LR), Random Forest (RF), and Support Vector Machine (SVM), and lodging datasets are utilized to look at the proficiency of calculations. Relative investigation of the classifiers shows that (LR) outflanks different strategies with high precision.

KEYWORDS: Machine Learning, LR, RF, SVM, algorithms.

1. INTRODUCTION:

AI (ML) is arranged under computerized reasoning of (AI) which works with the PC with proficiency to perform and learn even after not being especially modified. ML is a methodology for data assessment that robotizes sensible model structure. A current report from the McKinsey Global Institute declares that AI (a.k.a. information mining or farsighted assessment) will be the driver of going with the immense flood of headway.ML just focuses on creating PC programs adaptable to change at whatever point open to new information. Diverse ML calculations include the immense potential to be effectively applied in various fields like medicals, corporates, schooling, mechanical technology, games, and substantially more. In ML one of the significant variables is to make machines ready to adapt proficiently and adequately. There is a broad assortment of calculations that help in making contraptions and techniques in ML. At times these strategies make disarray in their relevance inappropriate techniques and which calculations give more precision can't know. Analysts have utilized various calculations in ML as per skill, accessibility, and the dataset Although ML is a tactfully youthful ground of exploration. Choosing calculations in ML for the given datasets (issue) can be interesting.

House value forecast is the most widely recognized investigation that we made in Machine Learning. Since it can deliver various outcomes on various calculations so it's smarter

to contract a few calculations with the test which one is better at foreseeing information with more exactness.

In this paper, USA-Housing Prices are anticipated utilizing Linear Regression (LR), Random Forest (RF), and SVM. The expectations over the cost of the house in the USA depend on a few things like level BHKS, area, Interiors, and so forth these calculations assist us with foreseeing more precise costs as per that. Cross approval is utilized with 70% for Train and 30% for Test the dataset.

2. THE USED REGRESSION:

2.1 Linear Regression (LR)

Linear regression [2] is an essential and generally utilized sort of prescient investigation. The general thought of Linear regression is to analyze two things:

(i) Does a bunch of indicator factors work effectively in anticipating a result (subordinate) variable?

(ii) Which factors specifically are huge indicators of the result variable, and how would they–demonstrated by the greatness and indication of the beta appraisals sway the result variable?

These regression gauges are utilized to clarify the connection between one dependent variable and at least one autonomous factor. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula:

$$y = c + b*x$$

Where,

y = estimated dependent variable score

c = constant

b = regression coefficient

x = score on the independentvariable.

There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressed. The independent variables can be called exogenous variables, predictor variables, or regressors. Three major uses for regression analysis are (i) Determining the strength of predictors

(ii) Forecasting an effect

(iii) Trend forecasting.

2.2 Support vector machines (SVM)

In ML, SVM is managed to learn models related to learning calculations that assess information utilized for classification and regression examination. In SVM planning estimation fosters a model that distributes new cases to one get-together or the other, making it a non-probabilistic equal direct classifier. At the point when realities are not classified, regulated learning is unimaginable, and a solo learning approach is mandatory, which endeavors to develop typical bunching of the information to gatherings, and afterward map new information to these framed gatherings. The clustering calculation which conveys an upgrade to the SVM [1] is called support vector grouping (SVC) and is frequently utilized in exchange applications either when realities are not ordered or when just a few realities are arranged as a pre-handling for a characterization pass.

The component of ordering the information into various classes by definition is a line that parts the preparation documents into classes. There are a couple of straight hyperplanes, SVM estimation attempts to increase the partition in the center of a couple of classes that are stunning and this is said as edge expansion. If the line benefits as much as possible from the space among the classes are perceived, the likelihood to work on well to unseen information is expanded.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

2.3 Random Forest

Arbitrary woods or irregular choice backwoods are a gathering learning technique [3] for classification, regression, and different assignments that works by developing a large number of choice trees at preparing time. For order assignments, the yield of the irregular woods is the class chosen by most trees. For relapse errands, the mean or normal forecast of the singular trees is returned. Arbitrary choice woodlands are right for choice trees' propensity for overfitting to their preparation set. Random backwoods by and large beat choice trees, yet their precision is lower than inclination helped trees. In any case, information attributes can influence their presentation.

The primary calculation for irregular choice timberlands was made in 1995 by Tin Kam Ho utilizing the arbitrary subspace technique, which, in Ho's detailing, is an approach to carry out the "stochastic separation" way to deal with the order proposed by Eugene Kleinberg.

An augmentation of the calculation was created by Leo Breiman and Adele Cutler, who enlisted "Arbitrary Forests" as a brand name in 2006 (starting in 2019, possessed by Minitab, Inc.). The augmentation consolidates Breiman's "sacking" thought and arbitrary determination of elements, presented first by Ho[1] and later freely by Amit and Geman to build an assortment of choice trees with controlled change.

Arbitrary timberlands are regularly utilized as "BlackBox" models in organizations, as they create sensible expectations across a wide scope of information while requiring little design.

3. EXPERIMENTAL DATASET AND METHODOLOGY:

In the current investigation of the dataset. Information is utilized USA Housing with 5000 lines and 7 segments.

The system for the relapse of the dataset is shown in figure 1. The examination has been acted in Jupyter journal running on (intel i5, eighth-gen) with 8GB RAM Installed.

3.1 Dataset:

The dataset has been downloaded from Kaggle.com.

3.2 Preprocessing:

After the social occasion of information, the following stage is to perform pre-preparing on the information. It is the procedure that changes the crude information to a justifiable arrangement.

3.3 Normalization:

It is a phase where every one of the qualities is changed to values somewhere in the range of 0 and 1. The justification for normalization is to draw in the data to a substitute scale.

3.4 ML calculations:

After standardization use, diverse ML calculations are utilized like Linear Regression, SVM, and Random Forest.

3.5 Validation:

In the approval, the cross-approval score strategy is utilized. To make the preparation and testing sets, all standardized elements are randomized before they can be utilized to prepare the relapse organizations. Direct relapse, Random timberland, and SVM are assessed. The partition is registered between test data and each instance of getting ready data. This division is chosen as the class of the test information.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

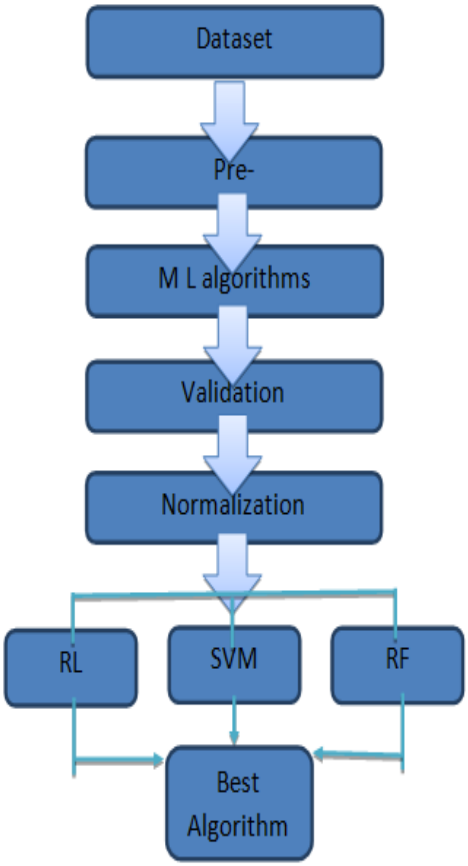


Figure 1: Schema of Experimentation methodology

4. RESULT AND DISCUSSION:

In the current work, the USA lodging dataset is utilized. To support and evaluate the execution of relapse, a cross-approval procedure [4] is used. The exactness of each crease is registered and the relapse models with the most elevated precision of the dataset are introduced in figure 2, and the histogram shows in figure 3. As indicated by figure 2 LR technique was utilized USA lodging dataset was applied on 70% train and 30% test. We got diverse astounding outcomes with a precision is 91.73%. Additionally, the Random Forest strategy was utilized where results were noticed the precision is 88.63%. For SVM the precision is 58.12%. Figure 2 demonstrated LR gives the best results conversely, with the others and it gives the higher accuracy with 91.73%.

Model	MAE	MSE	RMSE	RMSLE	R2 Square	Cross Validation
Linear Regression	94079.379374	1.410301e+10	118756.105679	11.684827	0.880493	0.917379
Random Forest Regressor	94273.167247	1.413682e+10	118898.773737	11.686028	0.880206	0.886343
SVM Regressor	87205.730510	1.172093e+10	108263.256765	11.592321	0.900679	0.581209

Figure 2: Accuracy Table

Figure 3 shows are the graphical portrayal [5] of the precision of applied strategies that is LR, SVM, and RF. Figure 3 (I) shows the R2 score matric to decide the precision of applied ML calculations, Figure 3 (ii) shows Mean Absolute Error (MAE) to decide the exactness of applied ML calculations, Figure 3 (iii) shows Mean Squared Error (MSE) to decide the precision of applied ML calculations, Figure 3 (iv) shows Root Mean Squared Error (RMSE) to decide the exactness of applied ML calculations, Figure 3 (v) shows Root Mean Squared Log Error (RMSLE) to decide the exactness of applied ML calculations, Figure 3 (vi) shows Cross-Validation Score to decide the exactness of applied ML calculations [6].

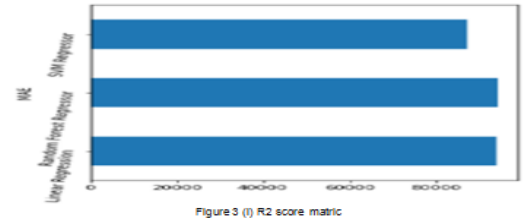


Figure 3 (I) R2 score matric

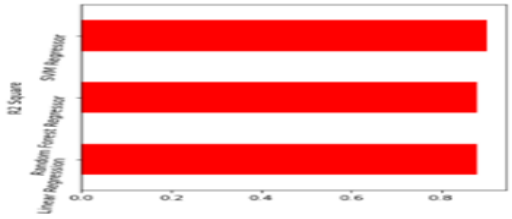


Figure 3 (II) Mean Absolute Error (MAE)

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

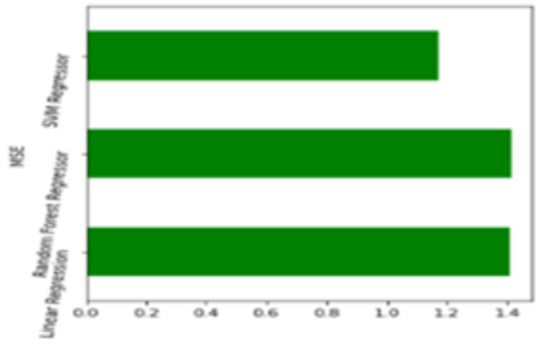


Figure 3 (iii) Mean Squared Error (MSE)

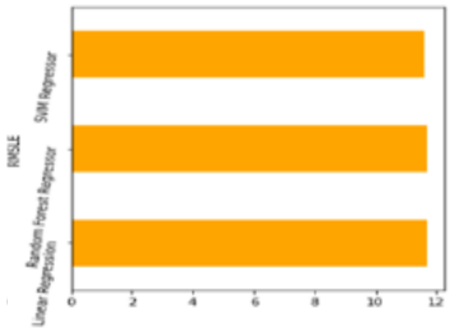


Figure 3 (iv) Root Mean Squared Error (RMSE)

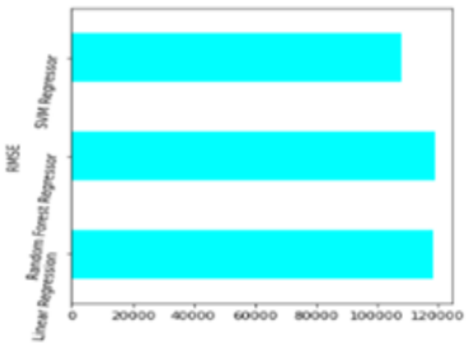


Figure 3 (v) Root Mean Squared Log Error (RMSLE)

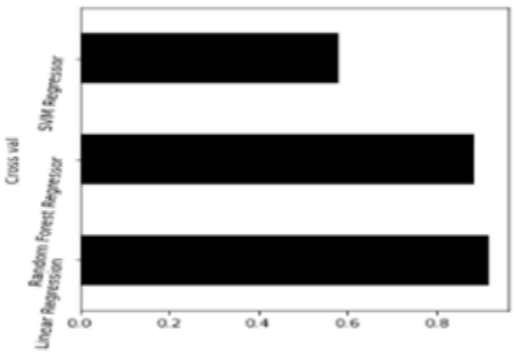


Figure 3 (vi) Cross Validation Score

Figure 3: Graphical Representation of Accuracy Metrics

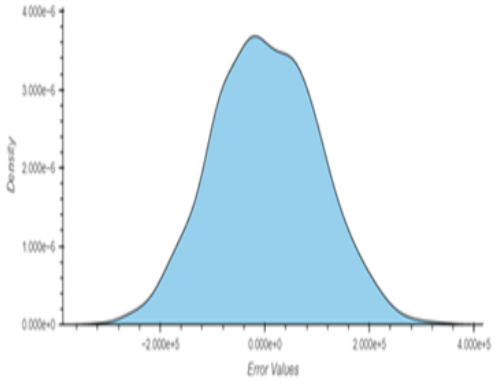


Figure 4 (i): KDE of Linear Regression

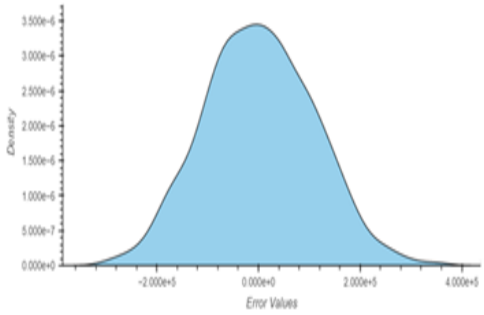


Figure 4 (ii): KDE of Support Vector Machine

Figure4 (iii): KDE of Random Forest

Figure 4: Graphical Representation of Error Value

5. CONCLUSION

According to the current assessment, the overall examination of LR, Random backwoods, and SVM has been executed and the execution metric mirrors the execution of LR better when diverged from the other relapse. The results surely exhibit the beginning that the features needed for the readiness of the relapse model should be solid and indisputable with the objective that various methods of administered learning can be researched to work on the execution. The exactness assessment of LR was seen to be near 91.73% it shows the higher adequacy of LR for anticipating. The LR-based models may be valuable in the field of determining, time-series Eexamination, forecasts, and various fields. According to the current assessment, the general examination of LR, Random woods, and

SVM has been executed and the execution metric mirrors the execution of LR better when appeared differently concerning the next relapse. The results surely exhibit the beginning that the features needed for the planning of the model for relapse should be solid and indisputable with the objective that various strategies of administered learning can be research to work on the execution. The exactness assessment of LR was seen to be near 91.73% it shows the higher adequacy of LR for foreseeing. The LR- based models may be valuable in the documentation of estimating, time-series, examination, forecasts, and various fields.

REFERENCES

1. F R Lumbanraja1, E Fitri , Ardiansyah , A Junaidi, Rizky Prabowo, "Abstract Classification Using Support Vector Machine Algorithm (Case Study: Abstract in a Computer Science Journal)", Journal of Physics: Conference Series (ICASMI 2020), 1751 (2021) 012042, doi:10.1088/1742-6596/1751/1/012042, pp. 1-12, 2020.
2. Shen Rong, Zhang Bao-wen, "The research of regression model in machine learning field", MATEC Web Conf., 6th International Forum on Industrial Design (IFID 2018), 176,01033(2018),https://doi.org/10.1051/matecconf/201817601033
3. Gerard Biau, "Analysis of a Random Forests Model", Journal of Machine Learning Research 13 (2012) 1063-1095.
4. Phan Thanh Noi, and Martin Kappas, "Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery", *Sensors* 2018, *18*(1), 18; https://doi.org/10.3390/s18010018.
5. Shengjia Cao, Yunhan Zeng, Shangru Yang, and Songlin Cao, "Research on Python Data Visualization Technology", Journal of Physics: Conference Series, Volume 1757, International Conference on Computer Big Data and Artificial Intelligence (ICCBDAI 2020) 24-25 October 2020, Changsha, China, 1757 (2021) 012122 IOP Publishing doi:10.1088/1742-6596/1757/1/012122.
6. Jin, H.; Stehman, S.V.; Mountrakis, G. Assessing the impact of training sample extraction on accuracy of an urban classification: A case study in Denver, Colorado. Int. J. Remote Sens. 2014, 35, 2067–2081.