2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) | 979-8-3503-9737-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACCS57279.2023.10113008

AutoML - Learning, Understanding and Applying Machine Learning to Datasets

Shreya Patankar
Department of Computer Engineering
K J Somaiya Institute of Technology
Mumbai, India.
shreya.patankar@somaiya.edu

Hitesh Prajapati
Department of Computer Engineering
K J Somaiya Institute of Technology
Mumbai, India.
hitesh.fp@somaiya.edu

Jeet Shah
Department of Computer Engineering
K J Somaiya Institute of Technology
Mumbai, India.
jeet15@somaiya.edu

Ankit Upadhyay
Department of Computer Engineering
K J Somaiya Institute of Technology
Mumbai, India.
ankit.upadhyay@somaiya.edu

***Abstract*—Current scenario of the digital world is loaded with sample amounts of data and to smartly analyze this data requires researching an interesting machine learning model to suit a dataset which is a topic of research and is more focused upon. Various state of art machine learning models like supervised, unsupervised and reinforcement models along with deep learning models are available to analyze varied datasets and for a given machine learning task, selecting one model from a large pool of options is a surprisingly challenging problem, particularly if there is a lack of evaluation data to distinguish between models and this requires for a thorough understanding of the dataset which you are using to solve the task. Dataset plays an important role in the performance of any machine learning model and not always you will get a clean / well balanced dataset to directly apply the model on. The key contribution of this paper is focused on the challenge of picking a good model which can be reformulated by doing a proper analysis and cleaning of the dataset, with the objective of recommending a suitable model for a given task based on the performance of a small set of probe models under the domain of classification and regression.**

Keywords—Classification, Regression, Dataset, Machine Learning, deep learning, Preprocessing, Recommendation system.

I. INTRODUCTION

Artificial intelligence (AI) being one of the biggest facets in this current digital scenario, it is used to detect patterns in data, identifying features, predicting models, analyze context to retrieve information and many more. The goal of AI is to make machines behave like humans. The major tasks performed by AI include prediction and regression analysis and classification. Prediction of weather forecast or disease detection, agriculture field prediction, remotely sensed data analysis and so on. Classification includes categorizing the data into groups based on the feature analysis. Performing any analysis requires to use the models of prediction, regression or classification and which model to use from the varied range of models is a difficult task and requires a thorough understanding of the dataset used in the process. In the current situation when one is asked to solve a problem for regression the user might not know what is the right model user should select according to the features of their dataset, many might find it difficult to clean their datasets for the optimal Machine learning result. Dataset plays an important role in the ML learning model as it results in performance of any ML task. Datasets if learned, understood and researched properly may help in determining the best suitable model for prediction, regression and classification problems.

In this paper we try to solve the above addressed problem and automate the entire process of cleaning the dataset, selecting the right model, and delivering high quality results to the user. Using our system, a person from a technical / non-technical background of machine learning would be able to learn, apply and understand how and why a particular machine learning model is preferred over other available models. We perform a deep study of all the state of the Art machine learning models and in depth study of all the available standard datasets to determine the best learning model.

In the domain of classification our system will recommend a model from Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees, Random Forest, Naive Bayes, Support vector Machine and in the domain of regression from Linear Regression, Logistic Regression, Polynomial Regression, Ridge Regression, Lasso Regression, Bayesian Linear Regression, Partial Least Squares Regression, Elastic Net Regression. The models are created by performing both classification and regression to build the model and analyze how well they are related to real world data in terms of accuracy, efficiency and predictability/recommendation.

The rest of the paper is organized as follows. Section I covers the Introduction of ML and the need of selecting the suitable data set. Section II covers the literature review on the ML models. Section III is focused on summarized findings from the Literature review, Section IV explores the objectives of the Research and Section V sums up the conclusion.

II. LITERATURE REVIEW

Zou et.al based on linear regression, specify the error function, determine the regression coefficients using gradient descent, and improve the sigmoid function to improve the accuracy of the binary classification. whereas the accuracy essentially stays the same, and the number of Iterations is decreased. The sigmoid function for bigger n, the number of iterations needed to attain It was determined to be, which is a modest number of times, after the binomial classification method was optimized. similar precision.[1]

Liu et.al in [2] use the logistic regression technique from the Sklearn machine learning package to categories a dataset of breast cancer (diagnostics). When we choose two characteristics with the highest maximum extent and maximum texture, the classification accuracy is 96.5%, which is an improvement over the prior technique.

Taunk et.al. in [3] demonstrated that the K-nearest-neighbor approach, which is straightforward yet extremely accurate, can be useful in some circumstances. This method has been used in the fields of healthcare and stock market forecasting, respectively. The KNN algorithm's K value is crucial since it influences the algorithm's precision and efficiency. Other extensions for the KNN algorithm have been suggested.Specifically, the SVM KNN classifier, the weighted KNN classifier, the shared closest neighbor KNN classifier, the locally adaptive KNN classifier, and the K-means KNN classifier. These increase accuracy and reduce execution time.

Navada et.al in [4] suggested that the IDA algorithm has a computational complexity that is half that of the ID3 algorithm. This method does, however, have significant shortcomings. Training data with empty attribute values cannot be handled; only a separate data collection may be. As a result, a superior algorithm was created that fixes all the problems of the C4.5 method. This method is more effective. takes care of repeated missing attribute values. It prevents overfitting and is resilient even when there is noise. also takes into account expenses for qualities. The ID3 algorithm does not take into consideration the interdependence of the variables. The overall categorization performance of the decision tree may suffer as a result. Additionally, the variables only have discrete values. Noisy recordings cannot be handled by this method. However, the ID3 algorithm generally performs a decent job at creating straightforward decision trees.

The Naive Bayes classifier discussed by Yang et.al in [5] employs probabilistic computations to ascertain the best-fit categorization for a given data set in a problem area. The Naive Bayes Classifier is more appropriate from the perspective of developing a generic toolkit for general classification expectations. Decision trees and support vector machines are typically used for binary classifications, but these two methods are restricted to having no more than two target classifications. They are difficult to generalize to meet a wide sense of real-life classification works because the number of target classes is typically greater than two.

The 20th century saw the development of kernels, which allowed SVM to be used for non-linear classification in addition to its original purpose of supporting linear classifications. SVM may be expanded to solve the issue when a set of samples cannot be linearly sorted into discrete categories. By employing kernel functions, the samples are mapped onto a high-dimensional feature space, enabling linear classification.[6]

Dichotomy-Rule-Fusion-Based Random Forest (DRF) was utilized in [7] for feature selection. It draws its inspiration from the concepts of information gain (IG) and recursive feature elimination (RFE). Using the ranking and impurity reduction provided by the features, it is utilized to choose the most significant characteristics. This enhances categorization accuracy.

For time series data, a comparison of linear regression versus support vector regression was performed and Kavitha et.al, in [8] used multivariate and time series data sets to analyze linear regression models utilizing the LeastMedSq and SMOreg functions. According to the analysis's findings, LeaseMedSq was the ideal model for linear regression.

Yang et.al in [9] demonstrated the convexity constraint of the loss function which is solved by using the filter theory from operations research and cybernetics to find the optimal parameters of the loss function. In this study, authors test the efficiency of the technique and apply logistic regression with filters to address the issue of real-world data categorization. By utilizing the filter theory from operations research and cybernetics, the convexity constraint of the loss function is resolved in this work in order to identify the ideal loss function parameters. In this study, we assess the effectiveness of the method and use logistic regression with filters to tackle the problem of classification of real-world data. The control objective is to ensure that the h-step output trajectory follows the planned reference trajectory. We can foresee an h-step future input sequence to achieve the control target thanks to model-free predictive control, which enables us to estimate the future input sequence from a major fraction of the recorded data.

Li et.al in [10] examined model-free predictive control for nonlinear systems using polynomial regression. Model-free predictive control doesn't use any mathematical models, yet it nonetheless achieves satisfactory control when large datasets are available close to the reference trajectory. It is therefore superior to the datasets that were previously maintained.

A generalization of polynomial regression, the Volterra series, a dynamic, nonlinear, time-invariant functional is functionally enlarged in mathematics by a Volterra series is employed in [11] to investigate model-free predictive control. Without having to estimate the polynomial regression coefficients, a suitable control input may be found using a dataset that contains the input/output data of the controlled system (Volterra series). As a result, it is critical to maintain a rich dataset, which means that the dataset must include input/output data that is close to the intended output. The dataset influences the control performance that is gained.

Literature survey concludes that it is simple to fit a variety of machine learning models on a given predictive

modeling dataset when using user-friendly machine learning tools like scikit-learn and Keras. The difficulty in applying machine learning then becomes how to select among a variety of models that you may employ to solve your problem. You could naively think that the model's performance is enough, but you need also take into account other factors, such as how long it takes to train the model or how simple it is to communicate with project stakeholders. The next section discusses the summarized findings from the survey.

III. Summarized Findings

We discovered that there was no comprehensive comparison or analysis for the best recommendation of a suitable model for the provided dataset after reviewing existing papers on Logistic Regression, KNN, Decision Tree, Naive Bayes, Support Vector Machine, Random Forest, Linear Regression, and Polynomial Regression. Additionally, for improved accuracy, it is necessary to clear out duplicate data, fill in missing values, and standardize the data value—steps that were not automated nor extensively covered in the articles.

Adding to the discussion of each algorithm's results, since it requires less iteration while maintaining the same level of precision, binary classification is a better fit for logistic regression, which is used for both regression and classification. There are several KNN versions, including the locally adaptive KNN classifier, the K-means KNN classifier, the weighted KNN classifier, the shared closest neighbor KNN classifier, and the SVM KNN classifier, all of which speed up processing and increase precision.

Missing values can be used with a decision tree. It resists overfitting and is noise-resistant. Most often used is ID3. In contrast to SVM and decision trees, which are appropriate for binary, Naive Bayes employs probabilistic computation. More than two classes can be classified using Naive Bayes, and it does so more effectively. SVM was developed to categories and address issues with non-linear classification. In order to classify features in high-dimensional feature space, it employs kernel functions.

Large datasets are more suited for Random Forest. Feature selection may not be necessary because it creates subsets of various features and applies models to them. All the models affect the total accuracy. It has a greater level of complexity and accuracy, and it may be used for both classification and regression. For regression issues when the characteristics are linearly associated, linear regression is performed. LeastMedSq is the superior linear regression model when compared to SMOreg. For polynomial regression to be effective with non-linear data, a large dataset is necessary.

The research surveys different ML models and their performance on various data sets. It gives us an insight about how different ML models performs and how each model extracts features and achieves satisfactory accuracy which helps us to understand the algorithms of ML and their performance on the data sets along with their accuracy. To the best of our knowledge and belief, there is very little research happening on the exhaustive study of Datasets which motivates us to take up this idea. In order to make the best use of resources and provide the same or a similar output for the given problem, we analyze multiple datasets on various models before coming to a decision on model selection based on accuracy and training time. Finally, for datasets that have not yet been seen, we would suggest models based on their features and content. Next section briefs the aims and objectives of our study.

IV. Objectives of the Study

The study aims to make the best use of resources and provide a same or similar output for a given problem by performing exhaustive study of datasets and analyzing them on various models before coming to a decision on model selection based on accuracy and training time. For datasets that have not yet been seen, we would suggest models based on their features and content.

V. Methodology

We start by evaluating several datasets and models, and then we build a rule-based system for model selection depending on the features and conclusions made from the dataset. Using the knowledge of model selection, for unseen datasets we begin with feature enrichment, selection, and data preparation before moving on to automatically deployed functioning machine learning models. Below mentioned Fig. 1 discusses the missing value analysis.

A. *Handling the missing values*

- Mean / Median / Mode
- Regression / Classification *algorithms as an imputer*

B. *Data Transformation*

- Label Encoding
- Normalization

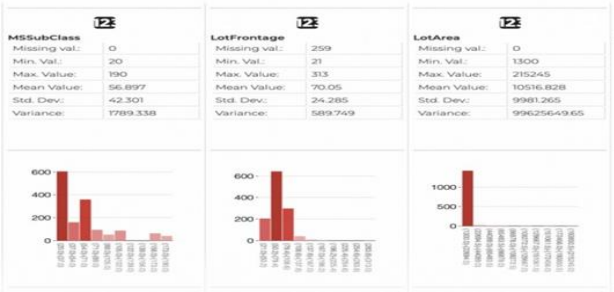

Fig. 1. Missing Value Analysis

VI. Conclusion

The study is more focused on understanding and applying ML techniques to varied datasets and extensively studying the dataset for recommendation to a suitable model. The datasets are analyzed to understand their features and how more suitable they tend to suit a model. The performance of the dataset is compared with state of the art which will be implemented as future scope.

References

[1] X. Zou, Y. Hu, Z. Tian and K. Shen, "Logistic Regression Model Optimization and Case Analysis," 2019 IEEE 7th International Conference on Computer Science and Network Technology

(ICCSNT), 2019, pp. 135-139, doi: 10.1109/ICCSNT47585.2019.8962457.

[2] L. Liu, "Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning," 2018 International Conference on Robots & Intelligent System (ICRIS), 2018, pp. 157-160, doi: 10.1109/ICRIS.2018.00049.

[3] K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbour Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.

[4] A. Navada, A. N. Ansari, S. Patil and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning," 2011 IEEE Control and System Graduate Research Colloquium, 2011, pp. 37-42, doi: 10.1109/ICSGRC.2011.5991826.

[5] F. -J. Yang, "An Implementation of Naive Bayes Classifier," 2018 International Conference on Computational Science and Computational

[6] S. Ghosh, A. Dasgupta and A. Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," 2019 International Conference on Intelligent Sustainable Systems (ICISS), 2019, pp. 24-28, doi: 10.1109/ISS1.2019.8908018.

[7] Y. Xiao, W. Huang and J. Wang, "A Random Forest Classification Algorithm Based on Dichotomy Rule Fusion," 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC), 2020, pp. 182-185

[8] Kavitha S, Varuna S and Ramya R, "A comparative analysis on linear Intelligence (CSCI), 2018, pp. 301-306, doi: 10.1109/CSCI46756.2018.00065. Conference on Green Engineering and Technologies (IC-GET), 2016, pp. 1-5, doi: 10.1109/GET.2016.7916627.

[9] Z. Yang and D. Li, "Application of Logistic Regression with Filter in Data Classification," 2019 Chinese Control Conference (CCC), 2019, pp. 3755-3759, doi: 10.23919/ChiCC.2019.8865281.

[10] H. Li and S. Yamamoto, "Polynomial regression based model-free predictive control for nonlinear systems," 2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), 2016, pp. 578-582, doi: 10.1109/SICE.2016.7749264.

[11] H. Li and S. Yamamoto, "A model-free predictive control method based on polynomial regression," 2016 SICE International Symposium on Control Systems (ISCS), 2016, pp. 1-6,