Effects of Feature Engineering on Machine Learning Model Performances

Aadyant Maity

Abstract:

This paper will investigate the effects of feature engineering on the accuracy of the Random Forest Regressor (RFR), Decision Tree Classifier (DTC), and Linear Regressor (LR) models when predicting the presence of a heart attack. By utilizing a tabular dataset of eight heart disease factors, we evaluate the models' accuracy when predicting a binary output relating to the presence of a heart attack. The findings highlight the remarkable potency of the DTC when predicting a binary value using tabular data points. They also highlight the detrimental effects on model accuracy of the incorrect utilization of feature engineering combinations. The valuable insights brought by feature engineering will contribute to the development of informed heart attack prevention measures because high-risk individuals can make informed decisions regarding their lifestyle with the help of accurate models.

Introduction:

With the prevalence of heart disease remaining a significant public health concern, the accurate prediction of heart attack susceptibility has garnered paramount attention. Heart attacks, a result of narrowing coronary arteries as a result of a build of fat, cholesterol, and other substances, are becoming increasingly common [1]. The gradual buildup in the arteries is called atherosclerosis [1]. Those with high blood pressure, blood cholesterol, and those who smoke are at extremely high risk for atherosclerosis and heart attacks [2]. In the realm of predictive modeling for critical medical conditions, this research centers on three prominent models - the Random Forest Regressor (RFR), Decision Tree Classifier (DTC), and Linear Regressor (LR) - in the context of predicting the presence of a heart attack. Drawing on a well-curated tabular

dataset encompassing eight distinct heart disease factors, the research meticulously assesses the models' accuracy in prognosticating a binary outcome related to heart attack occurrences. The implications of this research extend into the realm of practical medical applications. By unraveling the intricate relationships between model accuracy and feature engineering, this study equips healthcare practitioners and decision-makers with invaluable insights to empower high-risk individuals in making informed lifestyle choices. Ultimately, this research serves as a cornerstone in the edifice of evidence-based heart attack prevention strategies.

Methodology:

Dataset and Preprocessing

The dataset, comprising 1319 samples, serves as a valuable resource for investigating factors contributing to heart attacks, which account for a significant portion of CVD-related deaths. The dataset includes nine fields, encompassing demographic and physiological factors such as age, gender, heart rate, systolic and diastolic blood pressure, blood sugar levels, CK-MB and Test-Troponin levels. The primary objective is to elucidate the correlations between these attributes and the presence of a heart attack, classified into two categories: "0" indicating the absence of a heart attack, and "1" indicating its presence. Prior to analysis, a preprocessing pipeline was implemented, and involved data loading, attribute renaming, and binary encoding of the outcome variable for consistency. The resulting preprocessed data, organized as a NumPy array, forms the foundation for subsequent analyses and model development.

Feature Engineering Results

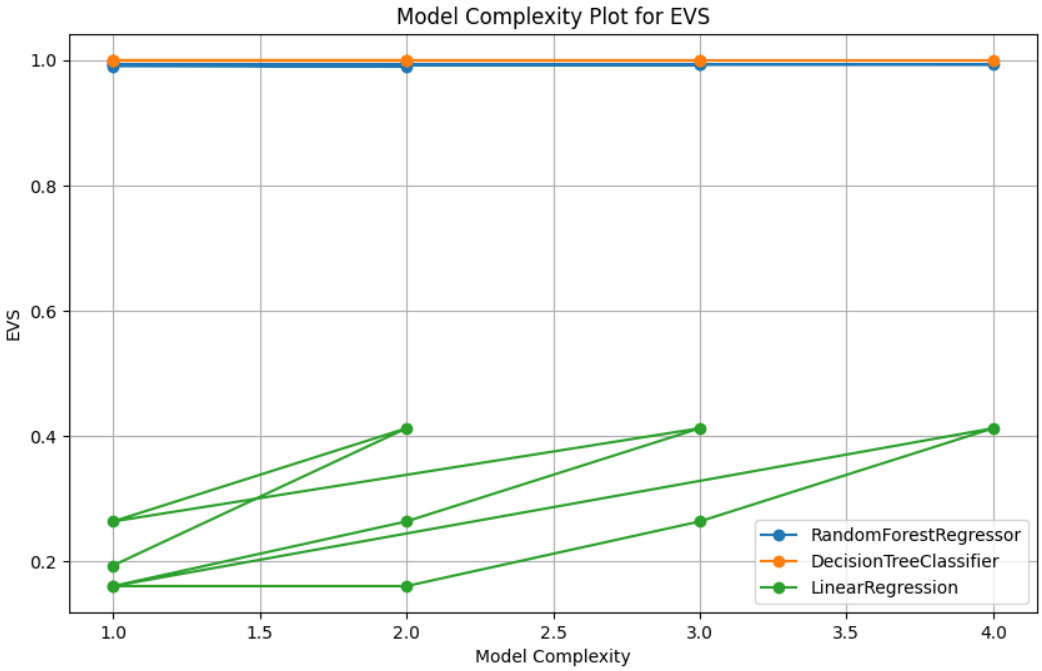
The data was used to train the models to effectively predict the outcome column based on the ten columns of data outlined above. The engineering was handled by the normalize(), generate_polynomial(), and generate_interactions() functions. The engineering showed us the

importance of utilizing an optimal combination of feature engineering techniques when it came to model accuracy by highlighting the loss in accuracy as a result of an inefficient combination of engineering techniques. I used normalization, feature interaction, and polynomial features to generate 9 sets of MSE, MBD, and EVS data. In the table below, the combo column represents the features that were used for the model evaluation. 0 represents normalization only, 2,3 represents feature interactions andpolynomial features only, and 2 represents feature interactions only. After reviewing the table, I discovered that DTC and RFR are not greated affected by a change in engineering combinations and were not of substantial help in this paper. However, the variability in accuracy for the LR model as a result of changing combinations is interesting and shows that feature engineering is crucial to the improvement of machine learning models' accuracies. This can be seen in the difference of 0.06 between the best and worst-case MSE scores for the LR model, 0.095 in MBD scores, and 0.262 in EVS scores. Being able to improve scores by at least 29.3% (MBD) represents the actual scale of how much feature engineering can improve machine learning models' accuracies.

Best				
Model	Combo	MSE	MBD	EVS
DTC	0	0.000	0.000	1.000
LR	2,3	0.139	0.324	0.412
RFR	2	0.001	0.006	0.994
Worst				
Model	Combo	MSE	MBD	EVS
DTC	0	0.000	0.000	1.000
LR	0	0.199	0.419	0.160
RFR	2,3	0.002	0.013	0.990

The table below shows how increasing the complexity (increasing amount of feature engineering) of the model affects the EVS score for DTC, LR, and RFR. One can see that

increasing the model complexity for DTC and RFR does not impact the EVS score to a large extent. Interestingly, the LR model's plot had a back and forth pattern that shows a potential drawback of feature engineering. This pattern shows that feature engineering combinations can be unreliable sometimes and can lead to degrading the models' accuracies rather than increasing them.



Conclusion:

In conclusion, this study delved into the effects of feature engineering on the predictive accuracy of machine learning models, specifically the Random Forest Regressor (RFR), Decision Tree Classifier (DTC), and Linear Regressor (LR), in the context of heart attack prediction. Through a meticulous analysis of a well-curated tabular dataset containing crucial heart disease factors, I sought to unravel the intricate relationship between feature engineering and model accuracy.

Our findings have illuminated several key insights. Firstly, the Decision Tree Classifier (DTC) emerged as a standout performer when it comes to predicting binary outcomes using tabular data, showcasing its remarkable potential in healthcare applications. Secondly, I underscored the significance of choosing the right combination of feature engineering techniques, as a suboptimal mix can adversely impact model accuracy. The stark differences in performance metrics, particularly for the Linear Regressor (LR) model, highlighted the crucial role of feature engineering in enhancing machine learning model accuracy. In the case of the LR model, feature engineering led to improvements of up to 29.3% in Mean Absolute Error (MBD) scores and 26.2% in Explained Variance Score (EVS) scores, underscoring the substantial potential for improvement. Additionally, the study demonstrated that increasing model complexity through feature engineering does not necessarily guarantee better performance, as seen in the LR model's fluctuating accuracy with changing feature combinations. This finding serves as a cautionary note, emphasizing the need for a thoughtful and data-driven approach to feature engineering. In the realm of practical applications, these findings hold great promise for healthcare practitioners and decision-makers. By understanding the nuances of feature engineering and its impact on model accuracy, high-risk individuals can make more informed lifestyle choices, potentially reducing their susceptibility to heart attacks.

Acknowledgements and References:

[1] - https://www.heart.org/en/health-topics/heart-attack/about-heart-attacks

[2] - https://www.cdc.gov/heartdisease/heart_attack.htm

Dataset - https://www.kaggle.com/datasets/bharath011/heart-disease-classification-dataset