Analysis of Regression Models on the Diabetes Data Set

Anonymous submission

Introduction

In this paper, we will be observing the effectiveness between different supervised regression models on the Diabetes dataset. This paper aims to help us to gain a sense of what factors are most strongly correlated to the cause of diabetes, identify which regression model gives the best interpretation of the diabetes data set, and what our metrics for a good model would be.

We will be testing between four different regression models: linear regression, kernel ridge regression, SGD regression, and LASSO regression. There are 10 different metrics or features that are included in the diabetes dataset and using these features, we will run a regression model to find which features are most likely correlated with diabetes.

The motivation to analyze the diabetes data set is to discover any patterns that are related between certain body indexes the diabetes dataset. If applicable, machine learning and its algorithms can be further applied to different medical cases besides diabetes and directly contributes to medical advances.

Related Work

Maulud and Abdulazeez (2020) reviewed the results of multiple papers on regression from 2016 to 2020. The different types of regression for the papers were single linear regression, multiple linear regression, and polynomial regression. The multiple linear regression papers were found to have the best accuracy. One of the papers they reviewed by Roopa and Asha (2019) used the diabetes dataset. They built a model to improve accuracy for the diabetes dataset. The model uses principal component analysis to extract features of the data and multiple linear regression to fit the model. The model was found to have 82.1 percent accuracy.

Fan, Chen, Li, and Zhu (2015) compared Lasso, adaptive lasso, ridge, and elastic net regression on the diabetes dataset. They look for variable selection and model prediction to evaluate the regression models. Akaike information criterion, bayes information criterion, and cross validation are used to evaluate variable selection. They find that lasso has the best variable selection, but has less accurate predictions, and elastic net finds the most accurate predictions.

Jayanthi, Babu, and Rao (2017) describe various predicative models for diabetes. They discuss models using neural networks, k neighbors, decision trees, support vector machines, and other classification models. There are also regression models like multiple linear, ridge, lasso, and elastic net. Hybrid models are also discussed, and they find the hybrid models to be more accurate. They find that elastic net would be the most useful for diabetes.

Methods

This paper will be covering four different regression models on the diabetes dataset: linear regression, kernel ridge regression, SGD regression, LASSO regression. Because the diabetes dataset has 10 different features to run regression on, each model was applied to the dataset 10 different times—each time for each feature. In total, 40 different iterations of the models were run on the same dataset to compare which model was the best applicable model for specifically the Diabetes dataset.

We collected data such as the coefficients, the mean squared error, and the coefficient of determination. Generally, the higher both types of coefficients were, the more likely the feature was in a correlative relationship with diabetes, and the lower the mean squared error was, the more likely the feature was correlated with diabetes. In addition, the average score for each of these models was calculated.

Results

As a result, it was found consistently across all four models that body mass index and s5, a measure of serum triglyceride levels, scored the highest coefficient values while maintaining some of the lowest mean squared error values. BMI scored the highest with triglyceride levels following behind. Average blood pressure (BP) was also was a close contender.

To compare between the models to determine which model would be the best one to model diabetes, we took an average of each score for every model, so an average coefficient, mean squared error, and coefficient of determination was generated for each model. From these models, it was observed that—by a small margin—linear regression was the best model followed by LASSO.

Discussion

Because BMI and triglyceride levels were the highest scoring features across all four models, it is highly likely that these features are correlated with diabetes. Although they are correlated, we cannot determine whether or not the features cause diabetes or whether diabetes causes the features to follow. Despite this, some useful information can be inferred. For instance, in a patient who does not have diabetes but has a high BMI or a measure of high serum triglyceride levels, it is likely that the patient may develop diabetes in the future.

The linear and LASSO models were also the best measures of the dataset, which implies that the data is linearly correlated; we can assume that a higher BMI and a higher recorded serum triglyceride level can be linked to a higher chance of someone having a case of diabetes. This chance changes linearly with BMI and serum triglyceride levels.

Conclusion

It was found that BMI and serum triglyceride levels were the most determining features in diabetes. The best models for analyzing the diabetes dataset in particular were the linear regression and the LASSO regression models. This was shown because of a combination of the lowest average error scores and highest coefficient values. BMI and serum triglyceride levels were found to be the highest scoring features across all of our regression models, reassuring their relevance with diabetes.

Some future work can include analysis on future state-of-the-art models to compare with the base models. LASSO is an instance of a quickly-evolving model that can applied in this script to compare the results with the previous iterations and/or other models. Currently, all the models would be considered under-fit due to the fact that none of the features for any models reaches a score over 0.5. This may be due to the limitations of linear regression models and other regression models such as quadratic models can be used.

References

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. Journal of Applied Science and Technology Trends, 1(4), 140-147. https://doi.org/10.38094/jastt1457

Roopa, H., & Asha, T. (2019). A Linear Model Based on Principal Component Analysis for Disease Prediction. IEEE Access, 7, 105314-105318.

Lei Fan, Shuai Chen, Qun Li & Zhouli Zhu (2015). ”Variable selection and model prediction based on Lasso, adaptive lasso and elastic net,” 4th International Conference on Computer Science and Network Technology (ICCSNT), pp. 579-583, doi: 10.1109/ICCSNT.2015.7490813.

Jayanthi, N., Babu, B.V. & Rao (2017). N.S. Survey on clinical prediction models for diabetes prediction. J Big Data 4, 26 https://doi.org/10.1186/s40537-017-0082-7