Regression Analysis of Medical Condition of Patient

Pranav Vinod

POM 681 - Business Analytics and Data Mining

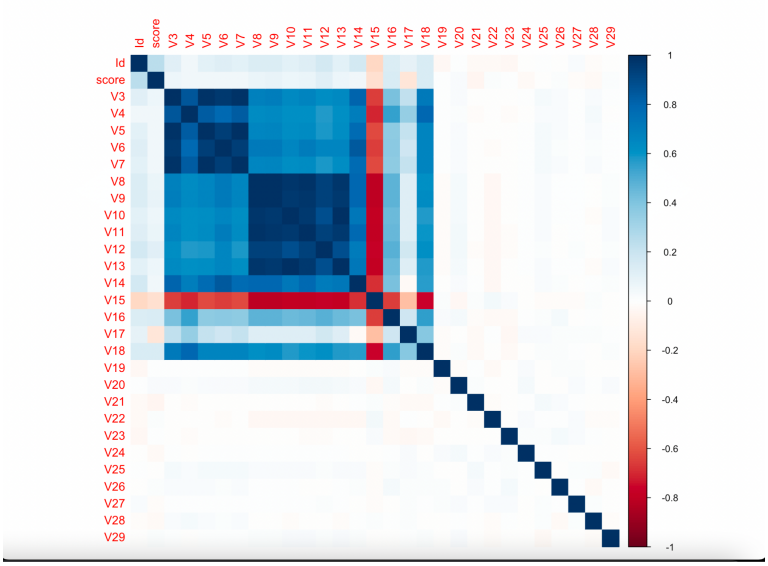
University of Massachusetts, Dartmouth

Abstract

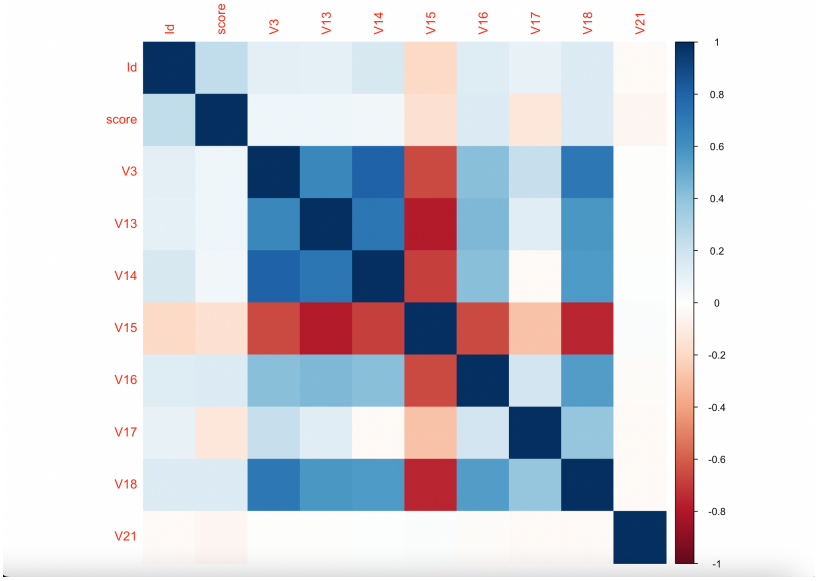
In this project, we have tried to estimate the score for a patient's medical condition using regression analysis. To achieve this goal, using exploratory data analysis, features about the dataset were looked into. Then the regression component of the project was implemented. Three models of regression were implemented to find the best model. Linear Regression, followed by random forest and XGBoost models were then implemented. Finally the results were analyzed to find the best possible result.

Exploratory Data Analysis

The dataset was part of Kaggle 2022 Regression Data Challenge with the aim of estimating score for a medical condition of a patient. The dataset consisted of 4141 observations of 29 variables. The target variable to be estimated was ‘score’. A summary of the data tells us that the data consisted of only numerical variables with no factors. The maximum score was 54.99 and the minimum was 7.00. On plotting pair plots of each variable against the target variable, we get to learn that despite being only a label, the column ‘Id’ had the highest correlation with the target variable. Further, on plotting the correlation plot of the dataset, we can notice that many variables are highly correlated with each other. Variables V3-V7 are highly positively correlated with each other and variables V8-V13 are also highly positively correlated to each other. This has the potential to skew our analysis and must be kept in mind. One more thing to take note of is the predictive power of the label ‘Id’. It seems to have the most correlation with our target variable which is unusual.



After dropping some of the highly correlated columns from the dataset, we obtained the following correlation heatmap:



After cleaning and engineering our dataset to get fewer variables, it was split into training and testing set in the ratio of 80:20.

Model

Initially, we implemented **Multiple Linear Regression** to the complete dataset. We found that the model was not a good fit for the available data. The model had a low R-squared value and a very low score on the F-statistic. But based on the multiple linear regression model, we were able to find the significant variables that are required to predict the target variable. We found that all the variables were significant. These were part of the table created after dropping the correlated variables.

Secondly, we implemented a **Random Forest** model to the dataset. The random forest updates a bagged decision tree model by choosing a random subset of the available features at each split. This ensures that all features have a chance to impact the final model. The prediction of each random forest is made by averaging over the predictions made by each tree.

Lastly, a **Gradient Boosted Decision Tree** Model was implemented. The boosted decision tree model updates the bagged decision tree model by using the outcome of each preceding tree to learn from and update the next tree. We found that the best results were achieved by optimizing the parameters of the gradient boosted model using an adaptive search and then fine tuning the parameters obtained.

Results

Three sets of results from the gradient boosted model are summarized in the table below:

	Without ID	Using ID	Using only ID
RMSE	8.69	3.00	1.99
Leaderboard Rank	86	46	27

The three sets of results were obtained using three different versions of the dataset:

1. Without ID - This result was obtained when the 'Id' column was removed from the dataset prior to model fitting.
2. Using ID -To investigate the effect of the column 'Id', this result was obtained when using the column 'Id' along with other features from the dataset. From this set, highly correlated variables were dropped prior to model fitting.
3. Using only ID - To further investigate the significance of 'Id', this result was obtained when fitting the target variable 'score' with only the column 'Id'.

We can see that the column 'Id' is the most significant when predicting the score of the medical condition of a patient. This would not be feasible in the real world as a patient id is not a reliable metric for predicting the score. This phenomenon when a variable from outside the scope of dependent features of the target variable is significant in predicting it is called data leakage in the dataset. There seems to be data leakage in this dataset.

Conclusion

On implementing three different regression models in this project, the best results were from the gradient boosted decision tree model. Then this model was implemented for 3 variations of the dataset and results were obtained. The best result was obtained when only the column 'Id' was used to predict the score of medical condition of the patient. This was found to be due to data leakage in the dataset. On excluding the 'Id' column, the model was found to have a Root Mean Square error of 8.69.

Reflective Statement

This competition has been a great learning curve for me. Having used datasets that required little exploration so far in my previous projects, working with a dataset that required me to analyze it before implementing machine learning models has been instructive. I was introduced to the concept of data leakage for the first time during the competition and it is something that I will keep in mind moving forward.

Further, I learnt about using adaptive search to find optimal parameters for decision tree models. Instead of manually fine tuning parameters, using adaptive search helped me investigate over a larger combination of parameters to find the optimal value for each parameter in the random forest and gradient boosted model. I will continue to refine my model and participate in the challenge.