Final Project Report

POM681 – Business Analytics & Data Mining
Evaluating the Medical Condition of a Patient

Prepared By: **Rohan Gonjari**

Student ID: **01987448**

Investigator: **Professor Bharatendra Rai**

University: **University of Massachusetts, Dartmouth, MA - 02747**

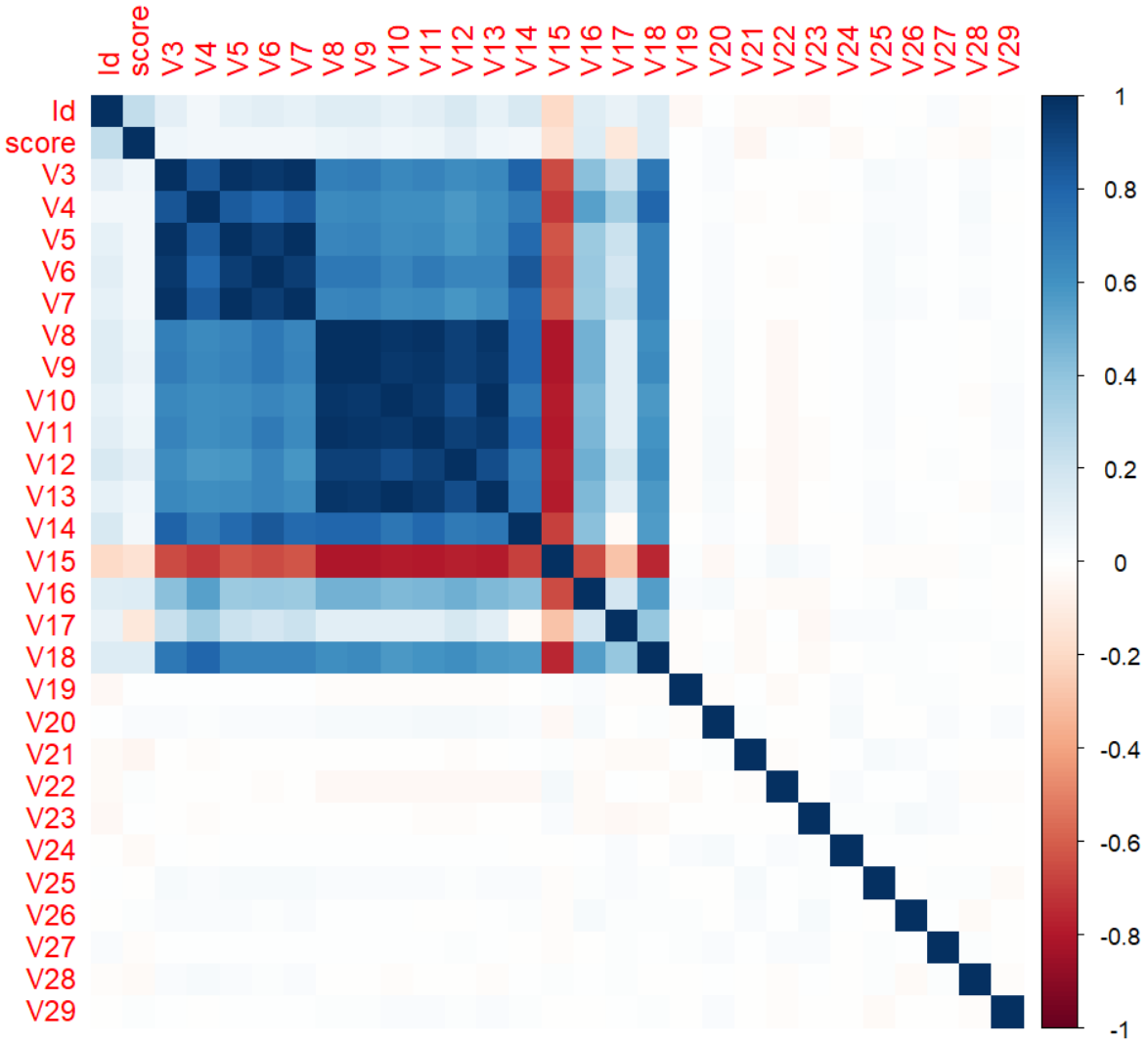
I. Abstract

The following project will aim to study the medical condition of a patient and predict medical scores using a model trained on the data provided. Initially, I worked on studying the dataset and trying to figure out which features are the most relevant to the medical score. I then proceeded to work with a few machine learning models to determine which model works best. The primary performance metric I used to evaluate the models is the Root Mean Squared Error (RMSE). Getting a lower RMSE also involved a lot of trial & error in trying to figure out which variables are significant. Once I had chosen the best working model, I then generated predictions for the test data provided.

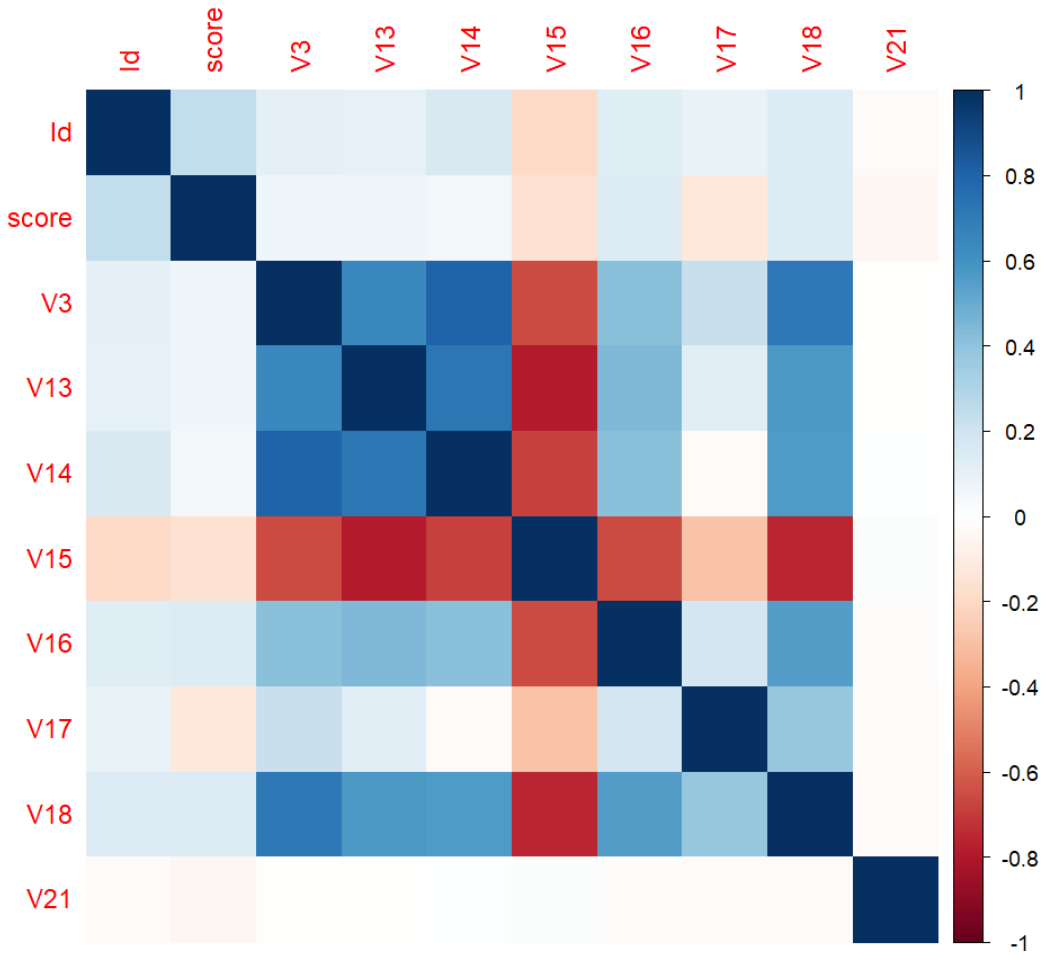
II. Data Preparation & Analysis

The training dataset was provided by the 2022 Regression Data Challenge on Kaggle. The dataset consists of 29 variables which include the target variable ‘score’. When studying the structure, we observe that the dataset consists of only numerical variables. The summary of the ‘score’ tells us the highest score is 54.99 and the lowest score is 7.00. I then proceeded to study the correlation plot for the dataset. The primary tool I used to study the dataset and the variables was the correlation plot. On plotting the correlation plot, we observe that many variables are highly correlated to each other. We note that variables V3 to V7 are highly positively correlated to each other and the same can be said for variables V8 to V13. Hence, I have only included one variable from both groups to minimalize any

skewness. Another strange observation we note is that the Id variable has a high correlation with the score. This indicates that Id can be a significant variable when predicting the score. However, this makes little sense when thinking from the perspective of a medical professional. Hence, we will further study why the Id has such a high correlation with the score. Below is the initial correlation plot we generated:



After getting rid of the insignificant variables and running a few trial multiple regression models to drop some specific variables, we generated our final correlation plot below:



For all the models below, I worked with two variations of the dataset, one with the Id and one without the Id, to further study how Id affects the model. I used a split of 75:25 for the training and testing sets.

III. Modelling

Multiple Regression:

I started off by building a multiple regression model with our selected variables. Once I has the model, I proceeded to generate predictions for the testing set and estimate the RMSE. The RMSE for the multiple regression was 10.29 without the Id and 9.94 with the Id. We were also able to confirm that the variables we had chosen were significant for our predictions. Given that there were much better RMSE estimates on the Kaggle leaderboard, I decided to move forward with other models.

Random Forest:

For our second model, I decided to employ the Random Forest algorithm. The random forest model consists of several decision trees. The model updates a tree by choosing a subset of variables from the available variables to help make a decision. The final prediction is the average of all predictions across all the trees. We had a fairly good reading when testing the model against the testing set. It generated an RMSE of 9.01 without the Id and 2.63 with the Id.

Gradient Boosted Decision Tree:

I decided to implement a Gradient Boosted Decision Tree given that in theory, it will outperform the Random Forest model. The Boosted Decision Tree updates the several decision trees in the model in a manner similar to the Random Forest model. However, it uses the outcome of the preceding tree to learn from and update its outcome. When the model was tested against the testing set, it generated an RMSE of 9.14 without the Id and 3.05 with the Id.

At the end of our modeling, I decided not to fine tune the Random Forest or the Gradient Boosted Decision Tree given that we had an acceptable accuracy with the relevant medical data and that the Gradient Boosted Decision Tree kept crashing my system.

IV. Results

I have compiled my reading into a tabular format below to help choose the best model:

Model	RMSE without Id	RMSE with Id
Multiple Regression	10.29	9.94
Random Forest	9.01	2.63
Gradient Boosted Decision Tree	9.14	3.05

V. Conclusion

My results indicated that the best-performing model I had was the Random Forest model. I then proceeded to test the Random Forest with the test data provided by the 2022 Regression Data Challenge on Kaggle. It generated an RMSE of 9.57 without the Id included in the test dataset. To improve my ranking further, I tested the Random Forest model with the Id included and generated an RMSE of 3.06. At the time of writing, my ranking is 60 on the leaderboard.

I noted how strange it was to include the Id throughout our entire project since it should not contain any relevance to the medical score. However, even when studying the correlation plot, we had observed how highly correlated the Id was to the score. When studying why the Id can be highly correlated to the target variable of a dataset, I came across a concept called data leakage. In essence, data leakage stated that there can be other hidden factors that highly positively or negatively correlate with the target variable and the Id too. This gives us the impression that the Id has some predictive power when it is the other hidden factors actually affecting the target variable.

VI. Reflective Statement

I initially faced troubles in trying to clean the data or decide which variables to drop or add. I only used the correlation plot as the primary tool to study correlation, so I did not have a lot to work with. However, I decided to employ a trial and error method using the regression models and confirm that the variables I had chosen were significant. I also learned how using a larger training set can lead to overfitting and can lead to a lower score overall. Therefore, I decided on a 75:25 split.

It was amusing generating predictions on the testing set and comparing them to the actual observations with and without the Id and observing how strange it was that the Id so strongly affected the score. This is where I was introduced to data leakage and how sometimes variables that seem irrelevant can affect the target variable.