Case Study Report

Introduction:

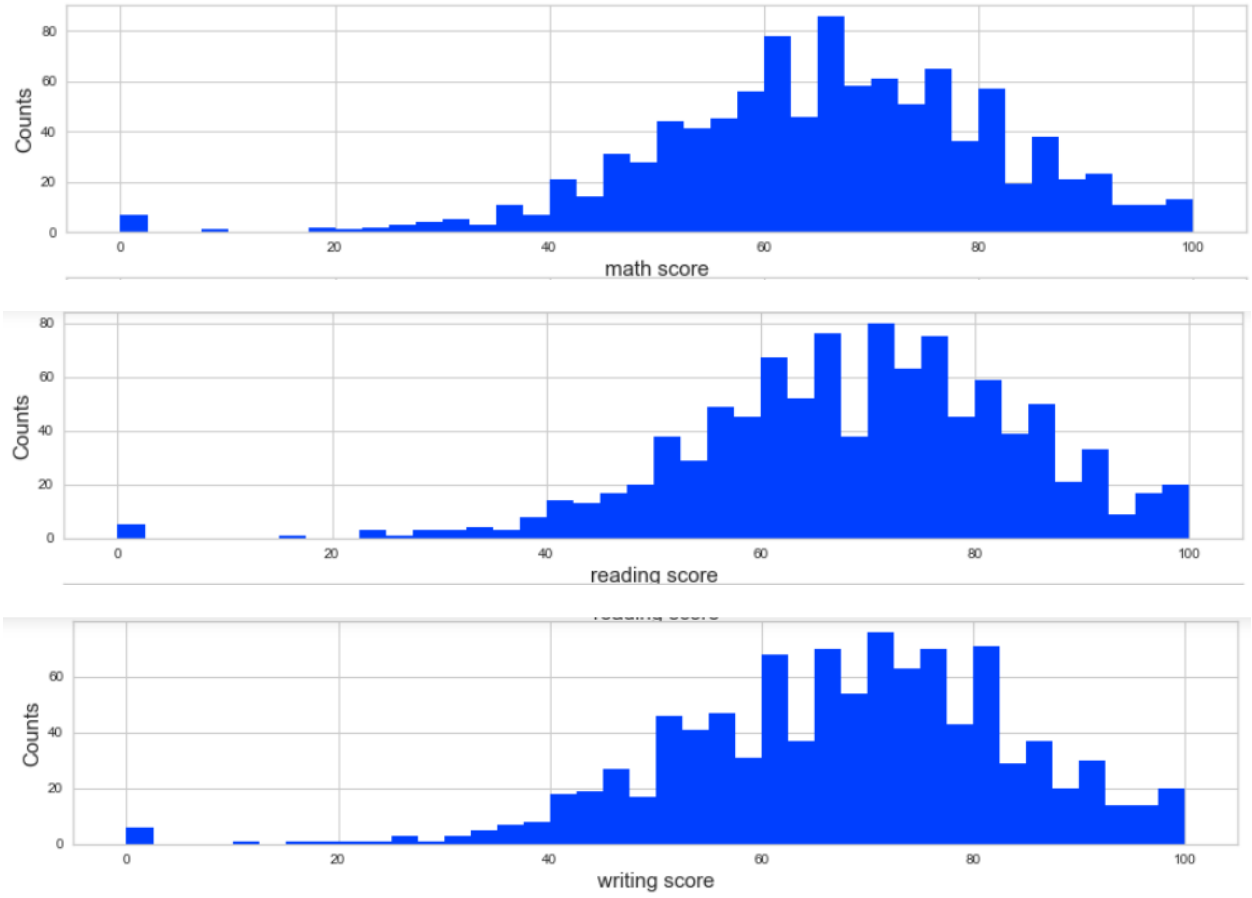
Logistic regression is one of the most fundamental and widely used Machine Learning Algorithm. A logistic regression in its plain form is used to model relationship between one or more predictor variables to a binary categorical target variable. The target variable is marked as 1 and 0. This case study is to predict the how the students performance which is the target variable based on the dependent variables of the marks they achieved in some subjects

Business Problem/Data:

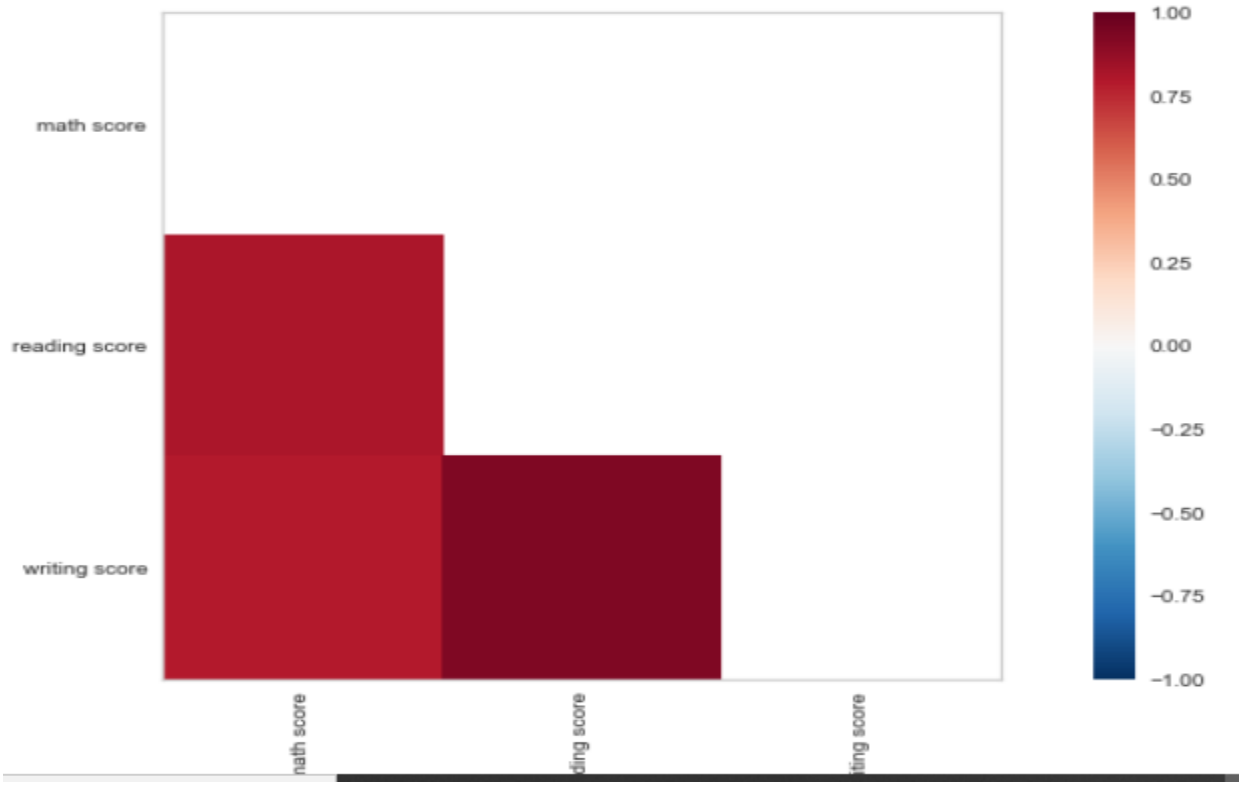
To build a simple logistic regression model for prediction of performance of the students given the values about their individuals scores they achieved based on which whether an entity can admit the student into their organization or not.

Graphical Analysis:

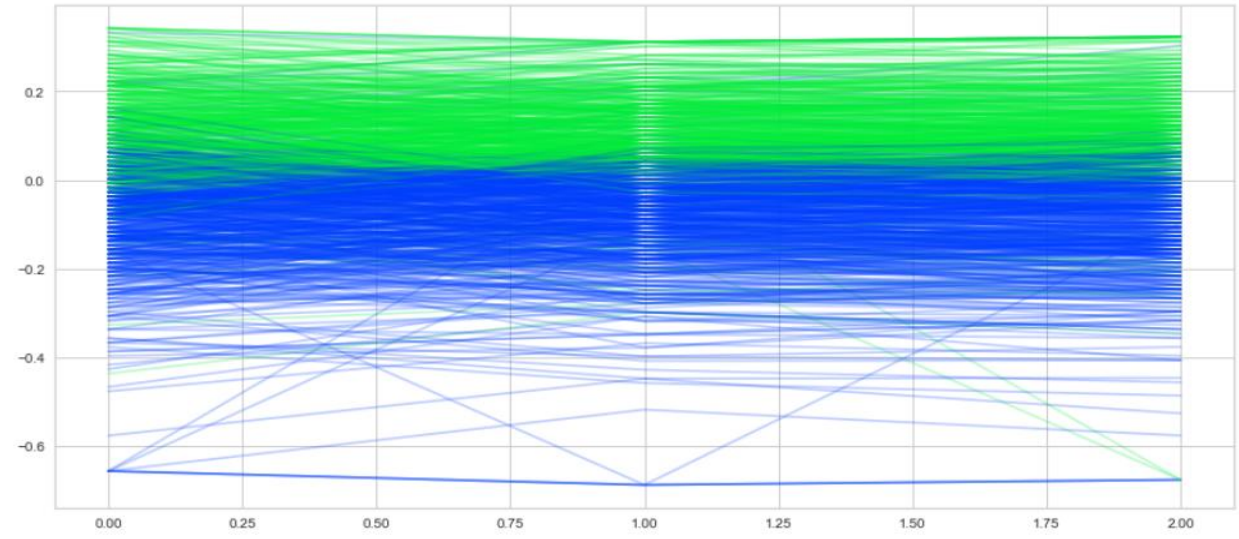
The below three bar graphs are a basic representation of the student scores in individual subjects



Below is the graph based on the pearson algorithm which explains us about the relationship between the variables chosen. Higher the coefficient, more dependency of the variables with one another. But our graph represents the less coefficient which indicates that these are not that much dependent on each other.



This graph represents the values of the selected features which were converted into vectors and plotted parallelly



Dimensionality & Feature Reduction and Feature Engineering:

Feature selection, also known as variable selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

Feature extraction is for creating a new, smaller set of features that still captures most of the useful information. This becomes even more important when the number of features are very large. There are many variables in the dataset which looked relevant to predict the performance of the student at first. After some thorough analysis was able to select the variables like reading score, math score and writing score to predict the outcome of the students performance.

The variables that we select for the building a model enables the algorithm to train faster, reduces the complexity and makes it easier to interpret. It improves the accuracy of a model if the right subset is chosen and reduces overfitting.

Model Selection & Evaluation:

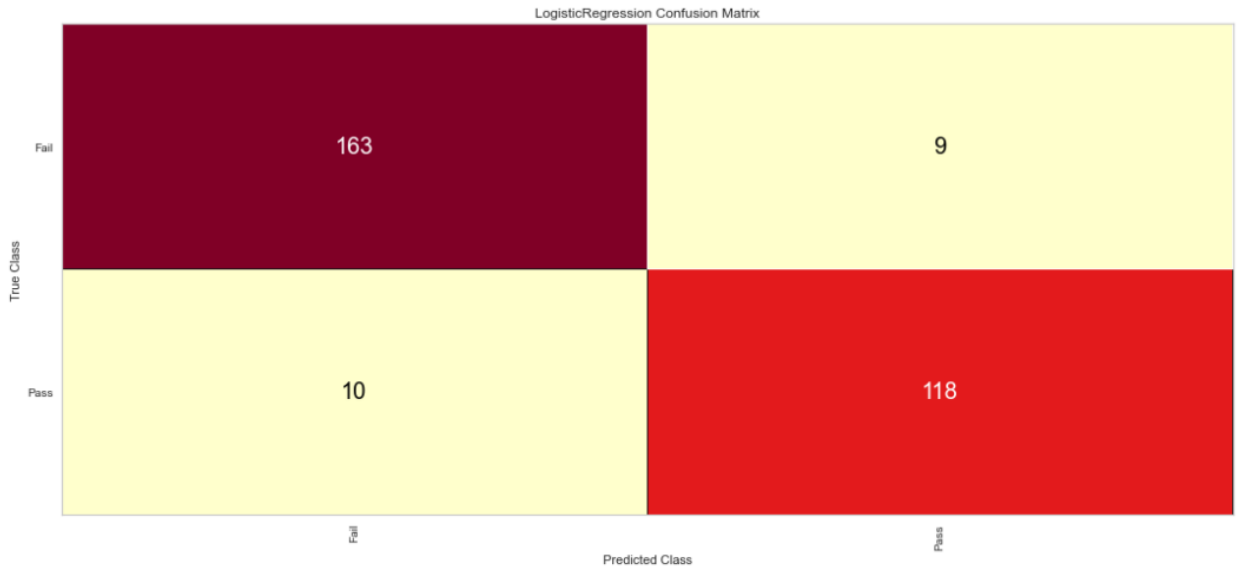
As the target variable that we chose for this is binary, I went ahead with Logistic regression to predict the outcome on whether a student is pass or not based on the scores in the individual modules. Below are the details of the training and validation dataset numbers

```
No. of samples in training set:  700
No. of samples in validation set: 300


No. of Pass and Fail in the training set:
0    380
1    320
Name: Pass, dtype: int64


No. of Pass and Fail in the validation set:
0    172
1    128
Name: Pass, dtype: int64
```

Accuracy of the Model: **93.67%**

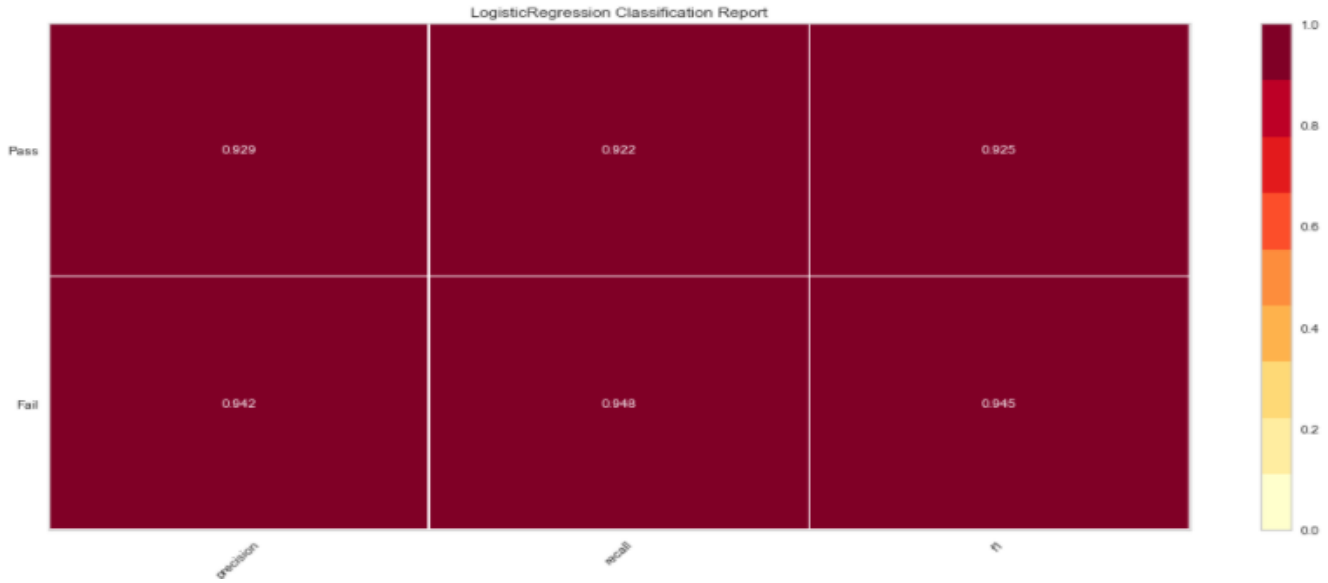


Based on the above confusion matrix

- TP = True Positives = 163
- TN = True Negatives = 118
- FP = False Positives = 9
- FN = False Negatives = 10

The left-hand side is the predicted values and the top the actual values. Just check where they intersect to see the number of predicted examples for any given class against the actual number of examples for that class.

Classification Report:



The term recall refers to the proportion of genuine positive examples that a predictive model has identified. To put that another way, it is the number of true positive examples divided by the total number of positive examples and false negatives.

Recall is the percentage of positive examples, from your entire set of positive examples, our model was able to identify. In our example it is **94.8%** which indicates a very good prediction.

Challenges:

I felt like the data that was used to perform the initial analysis should have been more selected more carefully. I could say it was the lack of knowledge because of which I missed a most important part of looking at all the variables in the dataset and check whether all the variables are in same type or not.

Conclusion:

Current model predicts that the scores are significant variables that decide the outcome of the students performance whether he/she passes or not. Also the accuracy of the model was 93.67% which implies a very good model with the data that we have at hand. To conclude higher the individual scores in each subject would increase the chance of passing in the semester. I can recommend this model to an educational institution whether to admit the student into their organization based on the marks they achieve.