Student Success Analysis

The Open University, March 2021

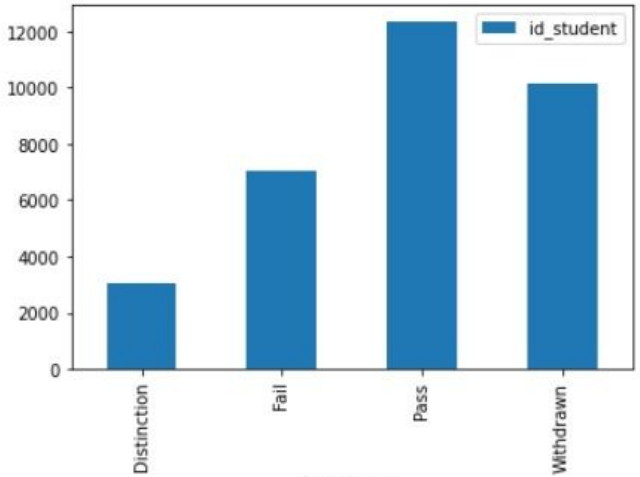
Suchitra Deekshitula

Agenda

Problem Area

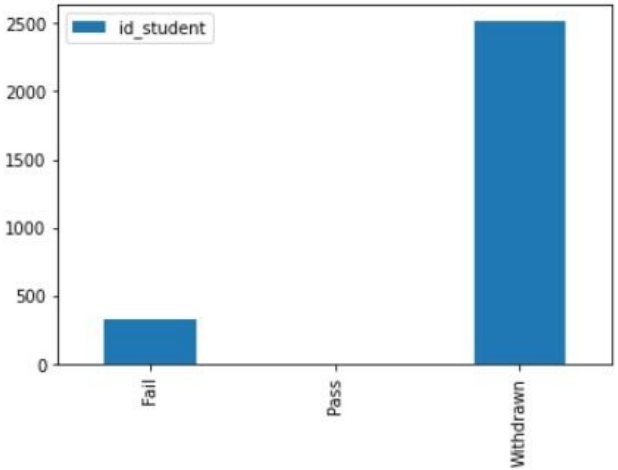
How can we improve pass rate of students? Does the Virtual Learning Environment (VLE) play a role in this?

Total Students in Courses



- We see that of the students enrolled in courses for the given presentations
  - 21.9% fail & 31.5% withdraw
  - <50% actually passed.

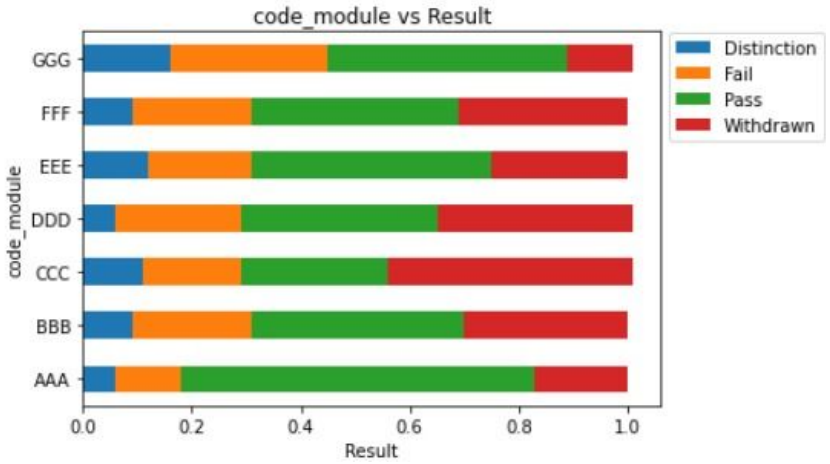
Students who didn't use VLE

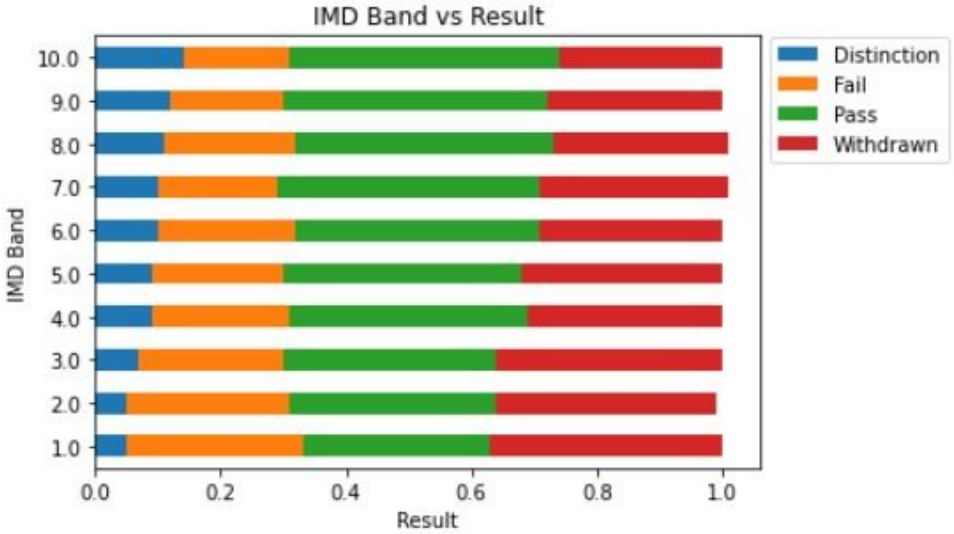


- 2711 Students! (9.4% of 28785)
- Students who do not interact with VLE have a higher chance of withdrawing and failing

This can be mitigated if we could identify students failing and withdrawing early in the program!

Result visualized by student attributes





Some code modules seem to have a higher fail and withdrawal rate than others

IMD band seemed to have some influence on the final result, more deprived the area, the more likely students are likely to withdraw and fail

[Result visualized by other student attributes here]

Analysis

- Using Gradient Boosting algorithm, we can **accurately predict results of 72% students**
- The model is likely to **deliver consistent results** on new data as it does not too closely fit the model to the limited set of data points used in this exercise (overfitting)
- The model does a **better job at predicting withdrawal** than fail

Detailed prediction percentages here

Glossary

Recommendations

- Since students who never interacted with VLE had a higher change of withdrawing and failing,
    - Adopting strategies to increase VLE engagement can have a positive impact in student success
- Considering some courses have a higher fail rate, additional due diligence on student past performance must be done before allotting courses to students
- Since students in less deprived areas (IMD Band) have a higher rate of failing and withdrawing, special focus towards them might help improve the results

Glossary

Future Work

- VLE interactions seem to have a significant impact on the result, we would need to collect more data to perform additional analysis

- We see that data for number of times a student interacts with the material (sum_clicks) is strongly correlated with the result

    - I would like to employ dimensionality reduction methods on this correlation in my future work to fine tune recommendations

Glossary

Appendix

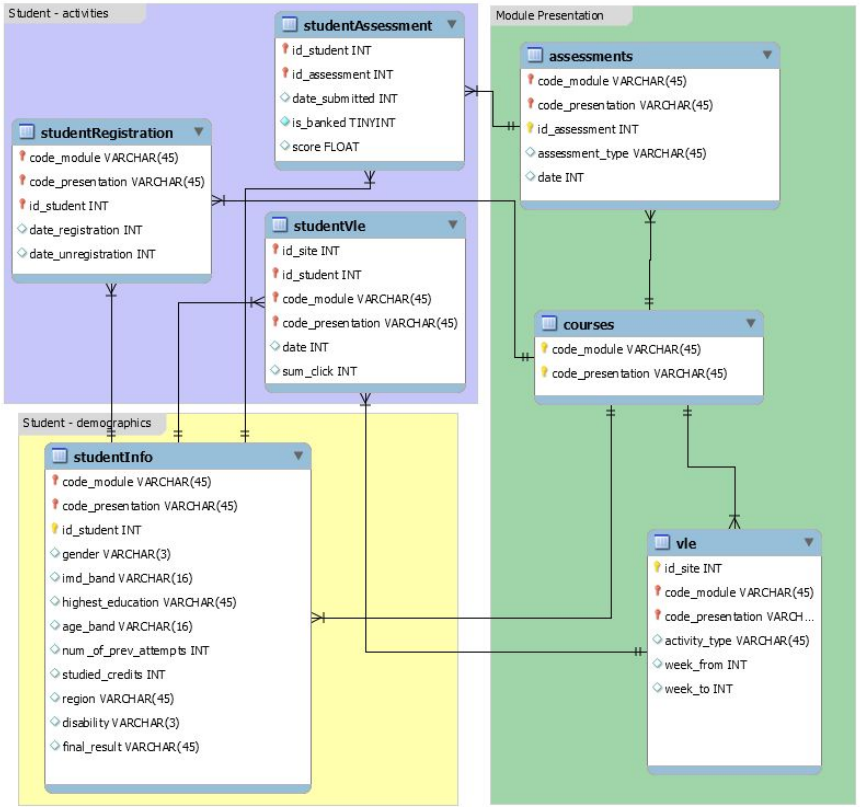
Data

Open University Learning Analytics dataset

- students in courses: 32953
- course-presentations: 22
- VLE pages: 6364
- VLE log entries: 10655280
- registration entries: 32953
- assessments: 206
- assessment entries: 173912

Glossary

Data: How are the datasets connected



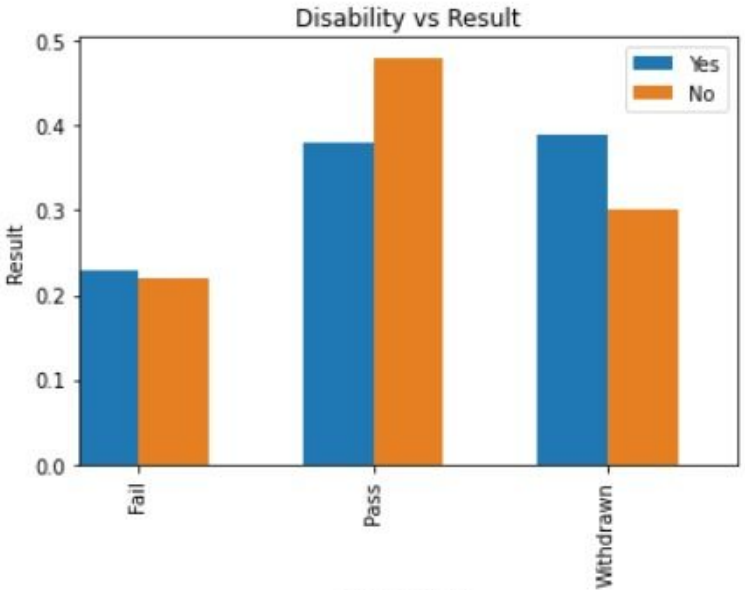
Glossary

Steps Followed

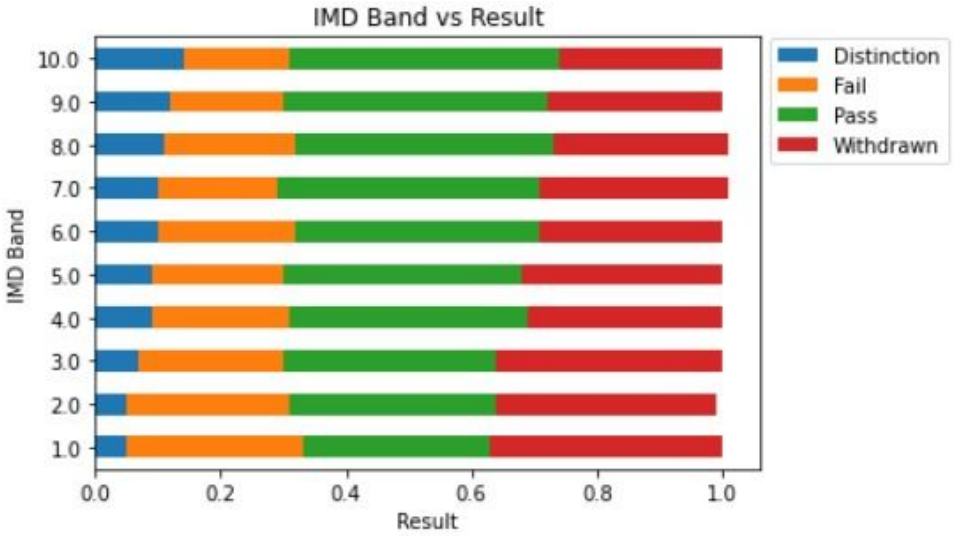
- Data Cleaning
- Exploratory Data Analysis
- Preparing Data for Modelling
- Predictive Modelling
- Results

Glossary

Exploratory Data Analysis



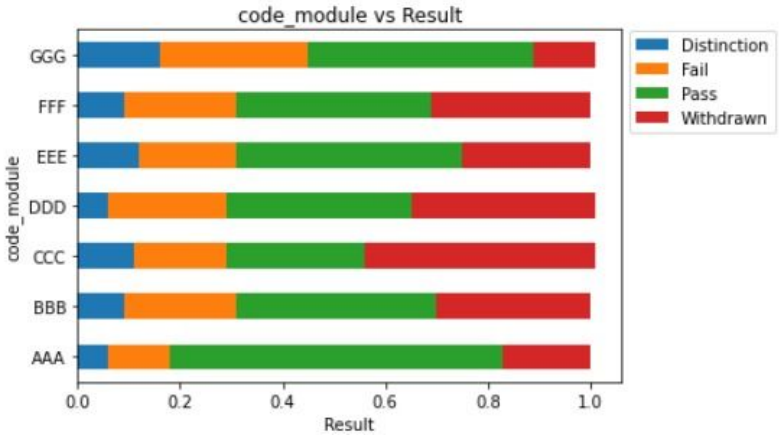
Disability status seemed to have some influence on the final result



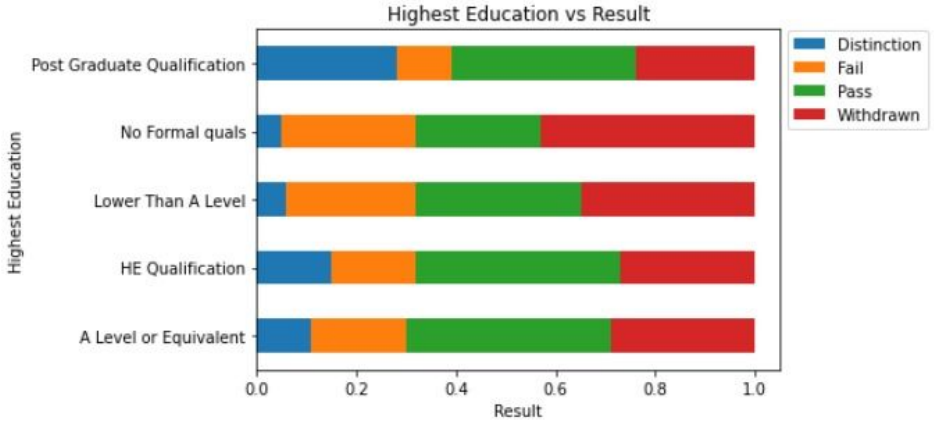
IMD band seemed to have some influence on the final result, more deprived the area, the more likely students are likely to withdraw and fail

Glossary

Exploratory Data Analysis



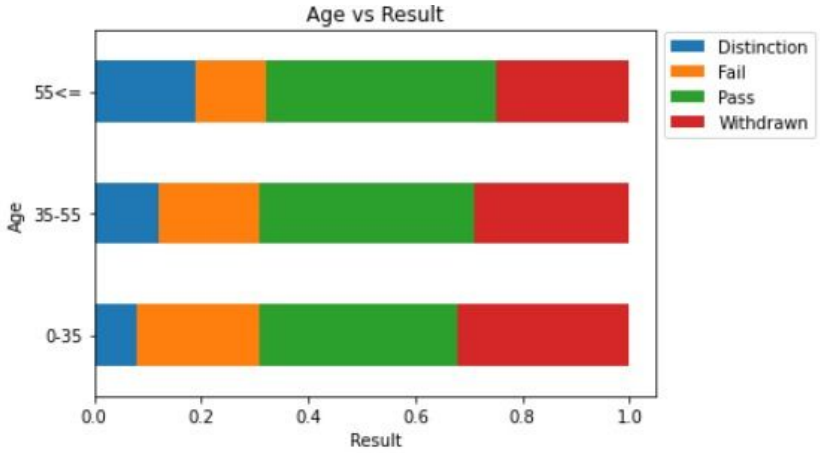
Some modules seem to have a higher fail and withdrawal rate than others



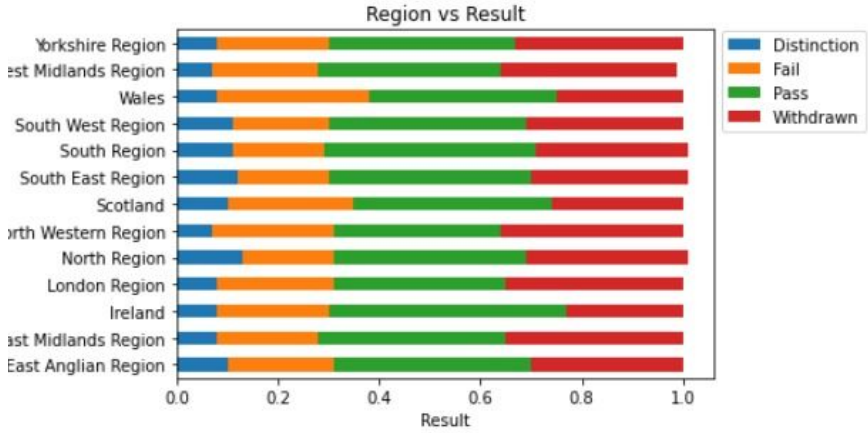
We see that students educational qualification with less educational qualifications have the highest fail and withdrawal rate

Glossary

Exploratory Data Analysis



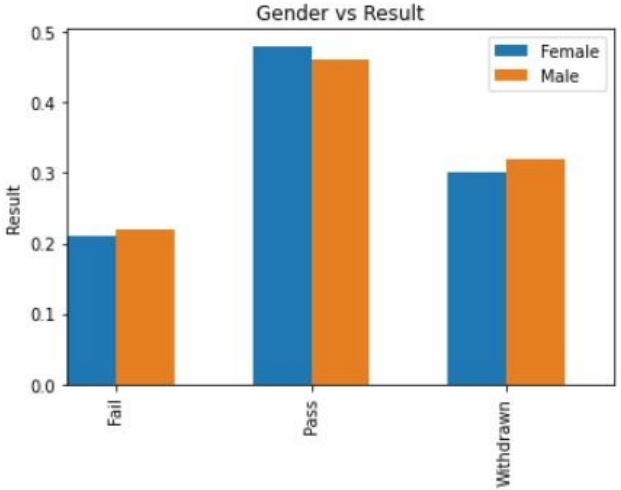
Age does not seem to have much of an influence on final result



Region does not seem to have much of an influence on final result

Glossary

Exploratory Data Analysis



Gender does not seem to have much of an influence on final result

Glossary

Preparing data for Modelling

- Most assessments happen towards the end of the course presentation, so we filter for assessments that happen before 100 days to predict early
- Understanding correlation between variables and final result
- Converting categorical columns to dummy variables
- Splitting data into training and testing datasets
- Upsampling data to balance classes

Glossary

Predictive Modelling

Decision Trees help create a model to predict the value/class of the target variable by learning simple decision rules inferred from prior data

- **Decision Tree:** The model is heavily overfitted and has a test accuracy of 63%
  - Testing Accuracy: 63%
  - Training score: 99%

Random Forests are an ensemble of many individual Decision Trees. Random Forest models combine the simplicity of Decision Trees with the flexibility and power of an ensemble model

- **Random Forest:** The model is heavily overfitted and has a test accuracy of 73%
  - Testing Accuracy: 73%
  - Training score: 99%

Glossary

Predictive Modelling

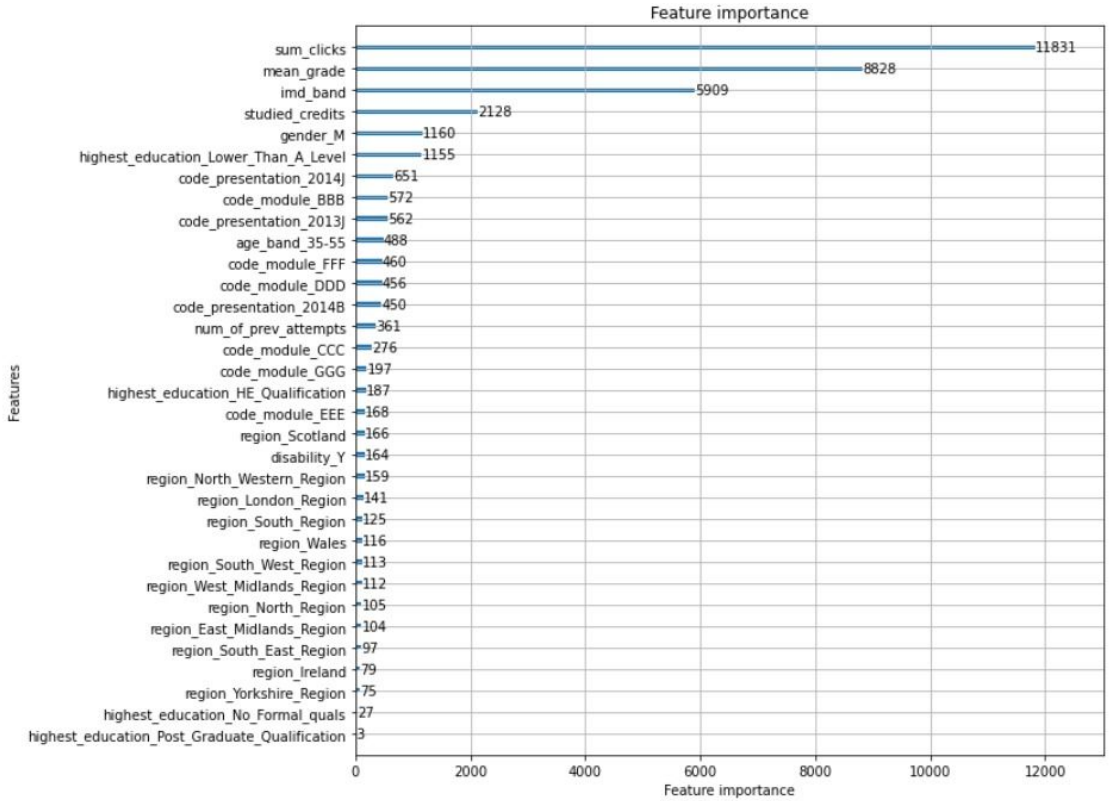
Gradient Boosting algorithm repetitively leverages the patterns in residuals and strengthen a model with weak predictions and make it better.

- **Gradient Boosting:** The model is slightly overfitted and has a test accuracy of 72%.
  - {'learning_rate': 0.2, 'max_depth': 80, 'min_child_weight': 40, 'n_estimators': 100, 'num_leaves': 160,'subsample': 0.6}
  - Testing Accuracy: 72%
  - Training score: 88%

This model suffers less from overfitting than decision tree and random forest. Hence, we will use this model for our recommendations.

Glossary

Feature Importance



The top 4 most important features are

- sum_clicks
- mean_grade
- imd_band
- studied_credits

Glossary

Model Conclusion

	precision	recall	f1-score	support
Fail	0.45	0.38	0.41	685
Pass	0.81	0.89	0.85	1576
Withdrawn	0.72	0.68	0.70	999
accuracy			0.72	3260
macro avg	0.66	0.65	0.65	3260
weighted avg	0.71	0.72	0.71	3260

0.7187116564417177

- We pick the Gradient Boosting algorithm with an accuracy of 72%
- The model predicts students who are more likely to pass.
- The model does a better job at predicting withdrawal better than fail

Glossary

Classification Report

- **Precision – What percent of your predictions were correct?**
  - Precision – Accuracy of positive predictions.
  - Precision = TP/(TP + FP)
- **Recall – What percent of the positive cases did you catch?**
  - Recall: Fraction of positives that were correctly identified.
  - Recall = TP/(TP+FN)
- **F1 score – What percent of positive predictions were correct?**
  - F1 Score = 2*(Recall * Precision) / (Recall + Precision)
- **The support is the number of samples of the true response that lie in that class**

Glossary

Classification Report Interpretation

- Percent of predictions that were correct (Precision)
  - Fail is 45%
  - Pass is 81%
  - Withdrawals is 72%
- Percent of positive predictions that were correct (F1 Score)
  - Fail is 41%
  - Pass is 85%
  - Withdrawals is 70%

Glossary

Glossary

- **Correlation:** Correlation analysis is a statistical method used to evaluate the strength of relationship between two quantitative variables
- **Overfitting:** Overfitting is a modeling error that occurs when a function is too closely fit to a limited set of data points
- **Decision Tree:** Decision Trees help create a model to predict the value/class of the target variable by learning simple decision rules inferred from prior data
- **Random Forest:** Random Forests are an ensemble of many individual Decision Trees. Random Forest models combine the simplicity of Decision Trees with the flexibility and power of an ensemble model
- **Gradient Boosting Algorithm:** Gradient Boosting algorithm repetitively leverages the patterns in residuals and strengthen a model with weak predictions and make it better.
- **Classification Report:** A Classification report is a tool used to measure the quality of predictions from a classification algorithm.
- **Dimensionality Reduction**: Dimensionality reduction is simply, the process of reducing the dimension of your feature set

Thank You