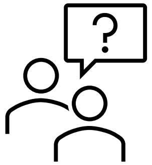
Classification and Features of Student Success in Online Programs

Jacob Javier

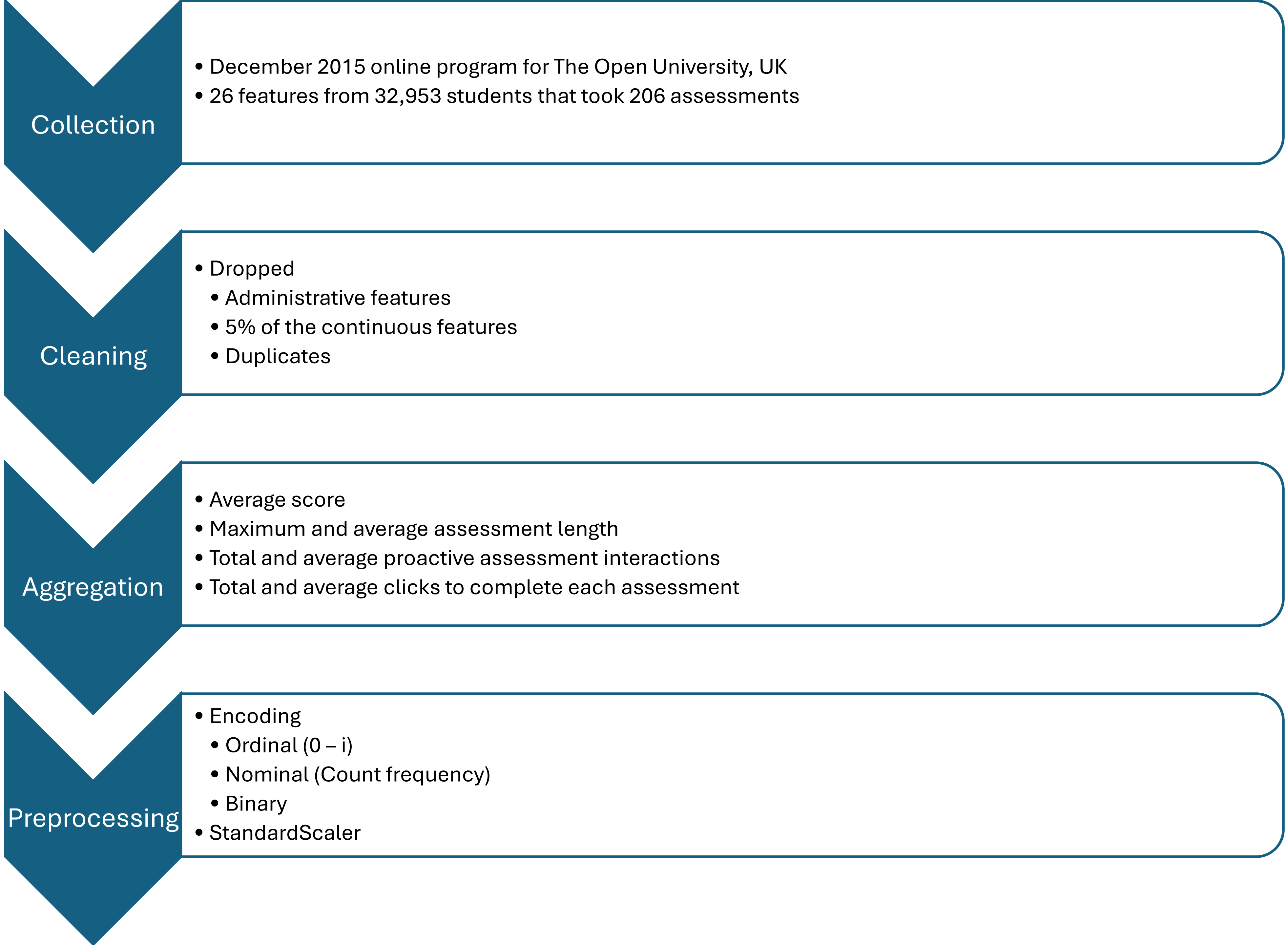
Springboard Data Science Career Track

Communication is one of the biggest challenges to online learning

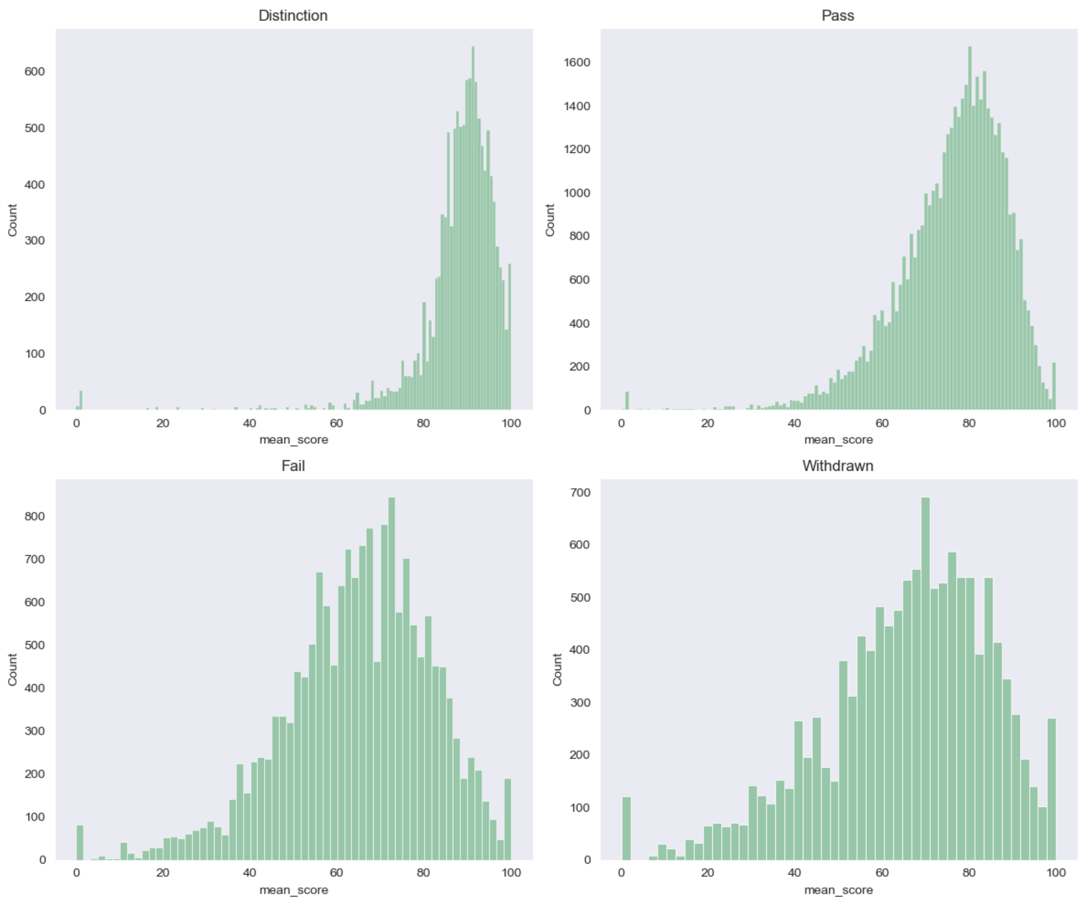


How can educators effectively gauge their students' engagement with the material in online programs?

- What features dictate student success in online programs, and can those features predict the trajectory of future cohorts as an advising tool?
  - The model must identify key features that determine final result classification.
  - The model must predict student outcomes to an 85% success rate.


Collection
• December 2015 online program for The Open University, UK
• 26 features from 32,953 students that took 206 assessments
Cleaning
• Dropped
• Administrative features
• 5% of the continuous features
• Duplicates
Aggregation
• Average score
• Maximum and average assessment length
• Total and average proactive assessment interactions
• Total and average clicks to complete each assessment
Preprocessing
• Encoding
• Ordinal (0 – i)
• Nominal (Count frequency)
• Binary
• StandardScaler

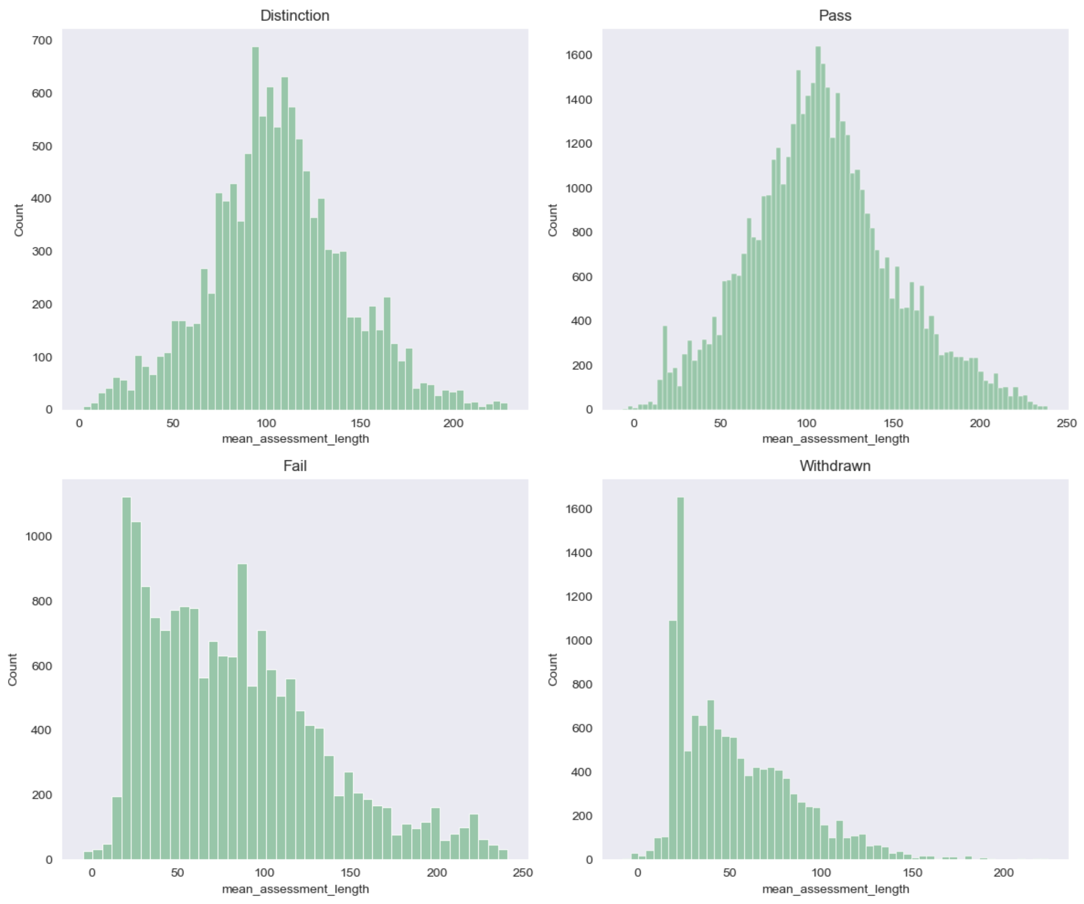
How does the average score for each assessment differ between the final results?



F-Stat	P-value
8332.65	0.0

Final Result	Average Score (%)
Distinction	89.28
Pass	77.03
Fail	65.15
Withdrawn	63.74

How does the material interaction for each assessment differ between the final results?



Stat	F-Stat	P-value
Average active	2.26	0.08
Average duration	6453.60	0.0

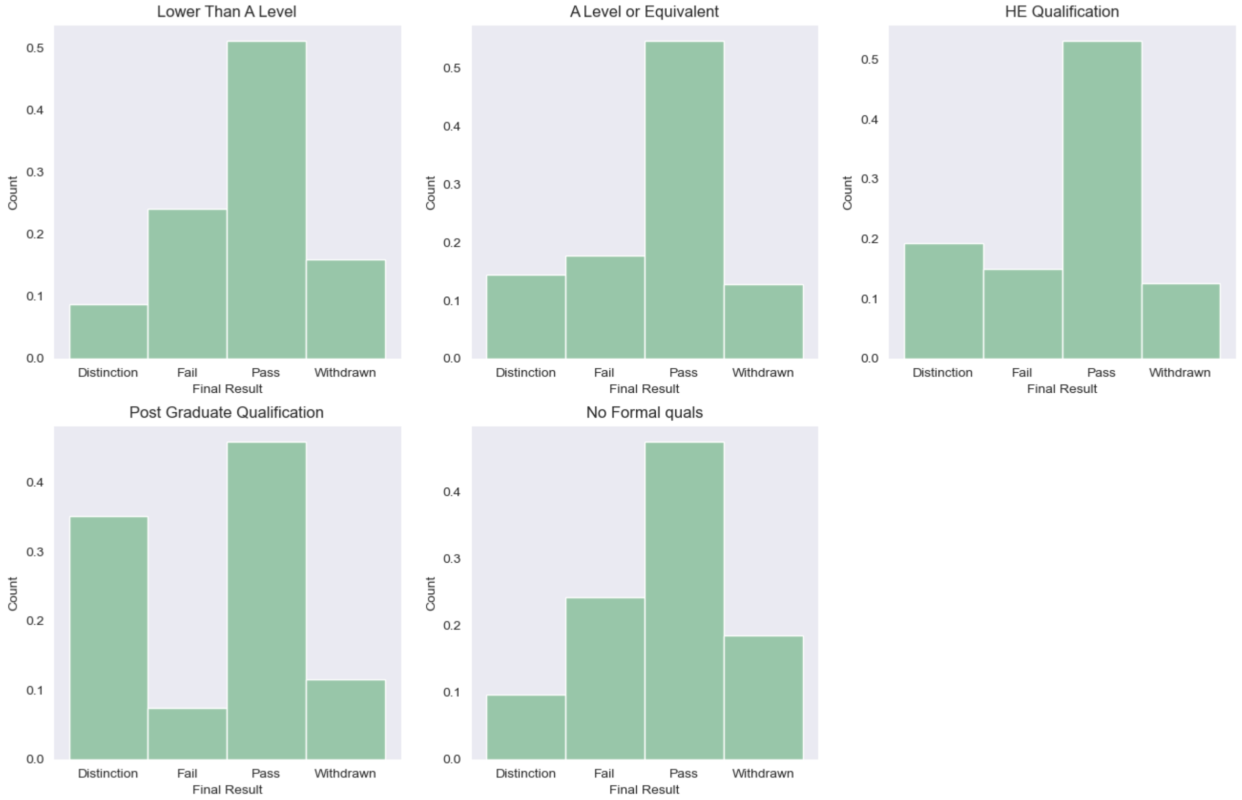
Final Result	Average duration (days)
Distinction	105.99
Pass	108.87
Fail	83.11
Withdrawn	53.93

Is there a difference between activity types that determine the final results?

χ^2	P-value
87.53	0.0

Is there a difference between activity types that determine the final results?

χ^2	P-value
26.22	0.0

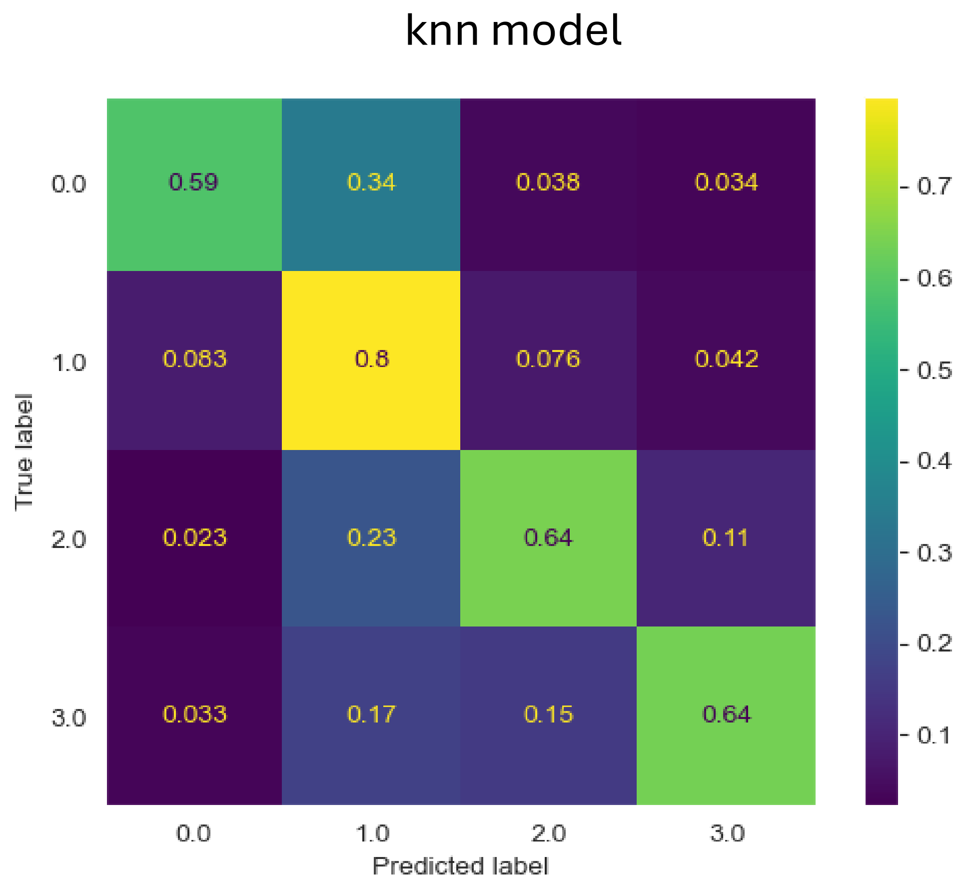


Supervised Multivariate Classification Modeling

- Splitting
  - 25% test set
  - Stratified: class imbalance
  - Shuffled: ordered data
- Hypertuning
  - RandomizedSearchCV
  - Stratified 5 fold
  - 250 iterations
  - Scoring: F1-score
- Models
  - Decision Tree (dt)
  - Random Forest (rf)*
  - K-Nearest Neighbors (knn)*
  - Logistic Regression (lr)

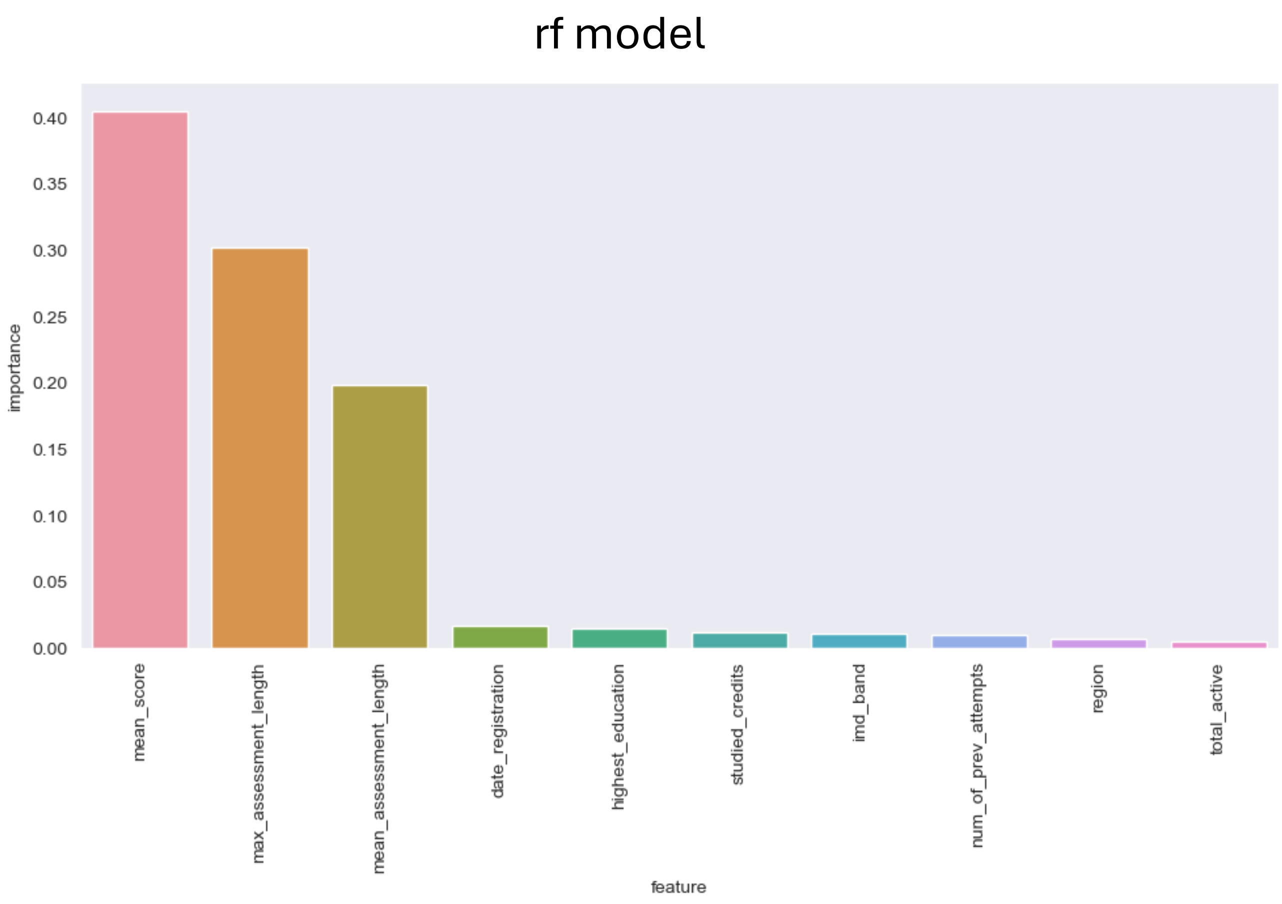
Classification

- K-Nearest Neighbors
  - Hyperparameters
    - Weights: Distance
    - Algorithm: Ball tree
    - Neighbors = 2
  - Metrics
    - Accuracy = 0.72
    - Weighted F1-score = 0.72
- Model best classified passing students
  - 79% of the students were identified correctly
  - Passing students account for 53% of the students



Feature Importance

- Random Forest
  - Hyperparameters
    - Max features: squareroot
    - Criterion: gini index
    - Estimators = 300
    - Max depth = 60
    - Min samples = 0.01
  - Metrics
    - Accuracy = 0.64
    - Weighted F1-score = 0.59
- Features
  - Average score (importance = 0.40)
  - Max Assessment Length (Importance = 0.30)
  - Mean Assessment Length (Importance = 0.20)



Conclusion

- The model must identify key features that determine final result classification.
  - Random forest identified average score and how long the assessments were active
  - The model had relatively mediocre metrics so new iterations will need to reassess feature importances
- The model must predict student outcomes to an 85% success rate.
  - The highest accuracy was 72% with the K-Nearest Neighbors classifier.
  - Using resampling techniques to reduce the class imbalance may improve the results.

Future Work

Data

- Rebalance classes
- Feature importances by
  - Average score
  - Highest education
  - VLE activity type

Additional Studies

- What factors lead to students turning in their assessments later than others?
- How has online learning changed between 2015 to present?