CAPSTONE PROJECT

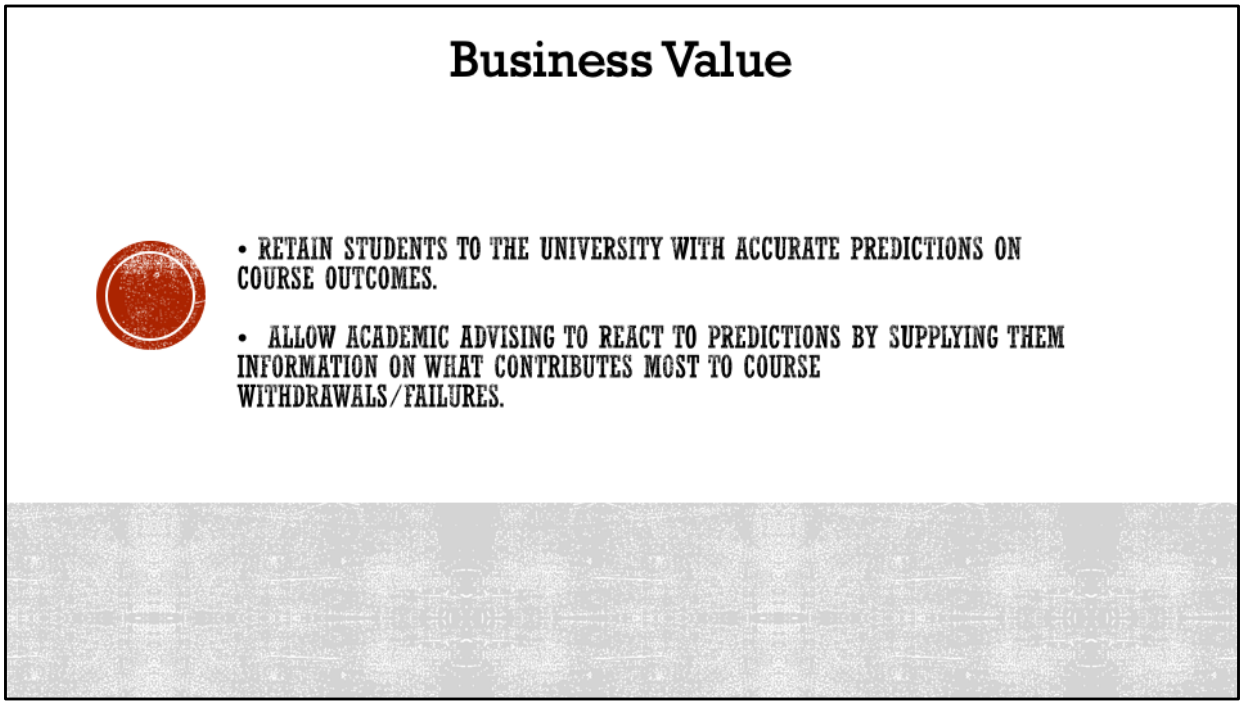
Michael Mathews Jr

Online-DS-PT-120919

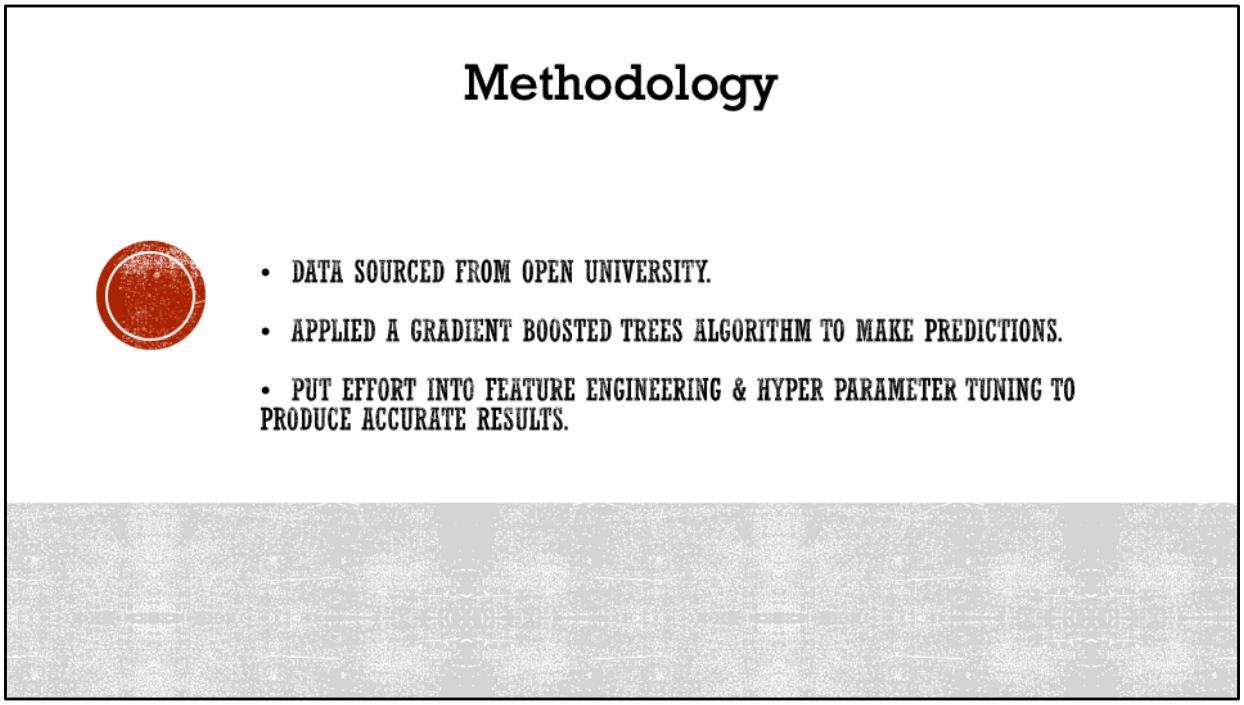
Problem Statement

ONLINE UNIVERSITIES HAVE A DIFFICULT TIME RETAINING STUDENTS. CAN I PREDICT COURSE OUTCOMES BEFORE THEY HAPPEN?

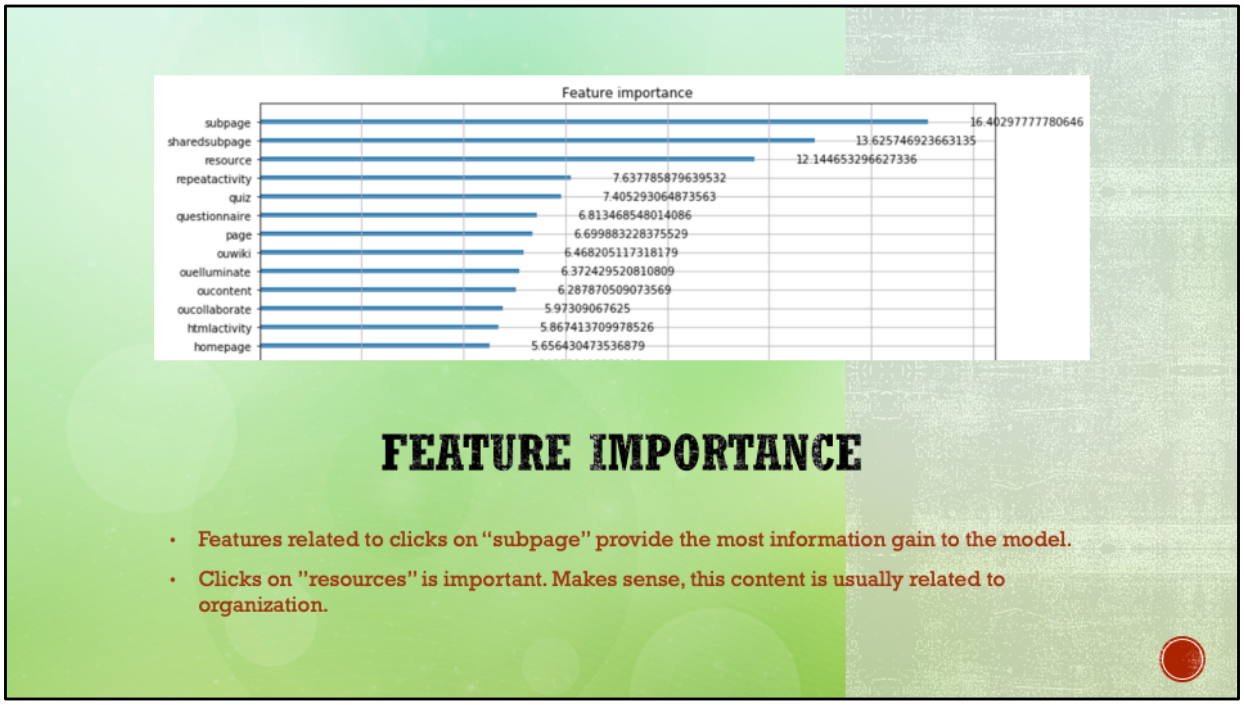
Can I predict which student courses are completed with a passing grade vs ones with a failing grade or withdrawal?



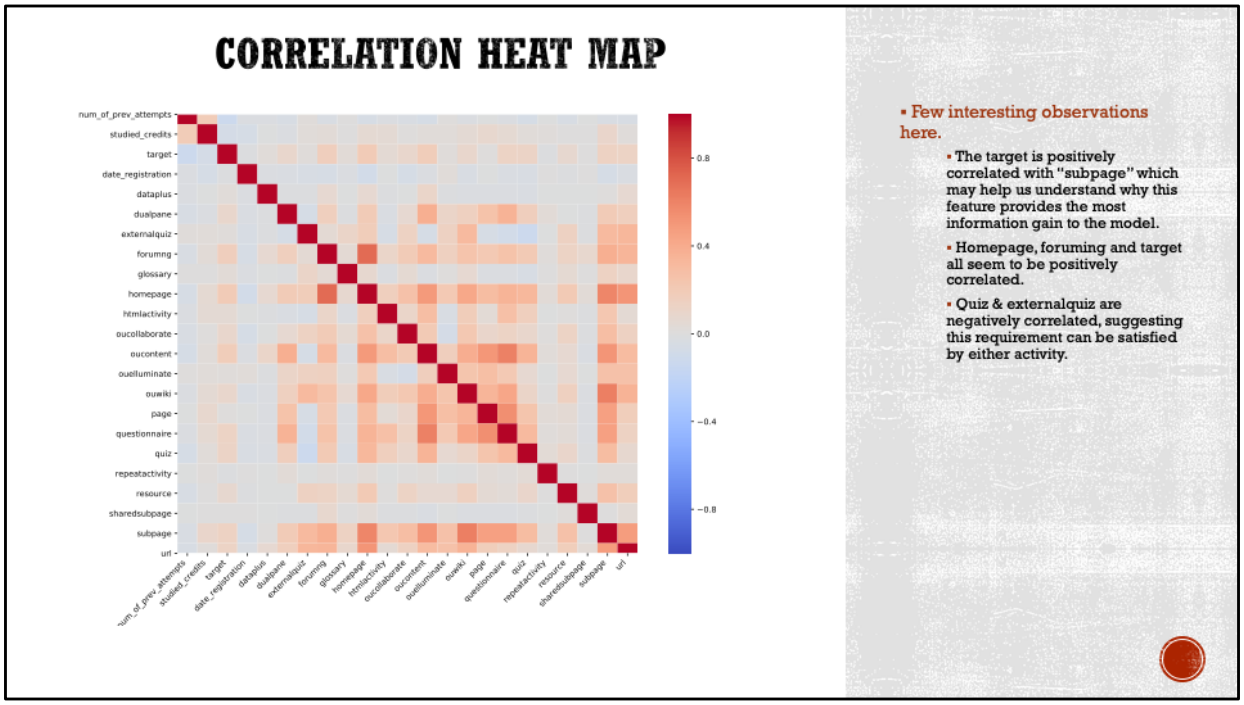
- The accuracy of the model should allow the business to leverage the information for financial gain as well as benefit student outcomes.
- Allocating time spent on advising students allows the university to waste little time. Instead of feeding “noisy” data with no statistical significance to the academic advising team, its better to model the course outcome using machine learning. Advisors will know which students to target their retention efforts.



- Data is sourced from Open University, the largest public undergraduate institution in the UK.
- I tested many algorithms, but settled on “XGBoost” because it gave me the most accurate predictions.
- Feature engineering is the most important part of the process. Extracting meaningful information out of the dataset that provides the most information gain to the model is most important.
- Hyper parameter tuning is less important with an algorithm like XGBoost because of its highly optimized state, but squeezing out more accuracy is always welcomed.

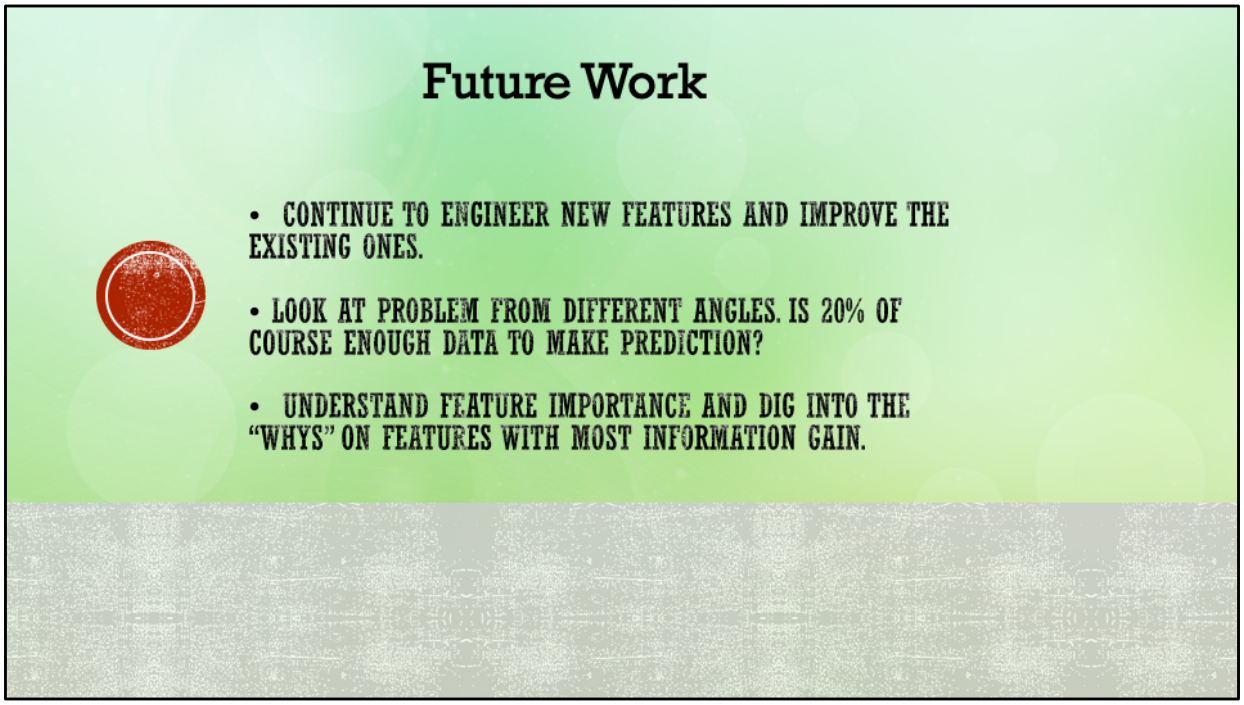


Plotting feature importance is a must when evaluating a machine learning model. The numerical scores you see are the total information gain supplied to the tree based classification model. I've taken note that the top two most important features are related to "subpage".



Heat maps are important to understand how the features are correlated with one another. It allows you to quickly scan each relationship. Both positive & negative correlations can uncover insights in the data set. This can lead to new or higher quality features. I would also say that features with no correlation at all can provide a data scientist with information.

- Its clear that evaluating a student course's data after 54 days of participation (20% of course length), provides valuable predictions that are accurate enough to act on.
- Evaluating millions of "clicks" on several virtual learning environment activities provides meaningful information gain to the model that produces 71.5% accuracy on course outcome predictions.



- Since feature engineering makes the most impact on predictions, it will be important to continue fine tuning them.
- Discovering new features makes it possible to raise accuracy a meaningful amount.
- Understand the domain better by exploring the features that provide the most information gain. What is it about these features that are informing the model? This can help with engineering new features.

- Questions, comments and/or feedback? Thank you for following along!