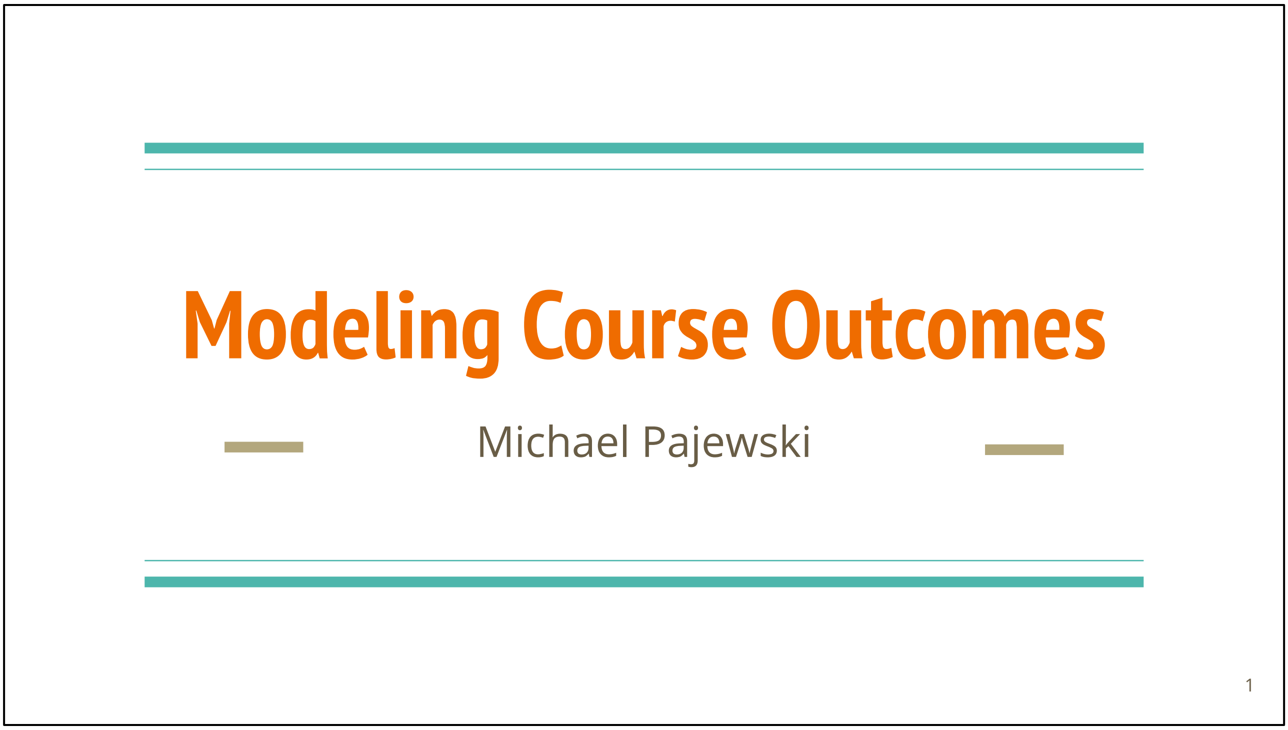
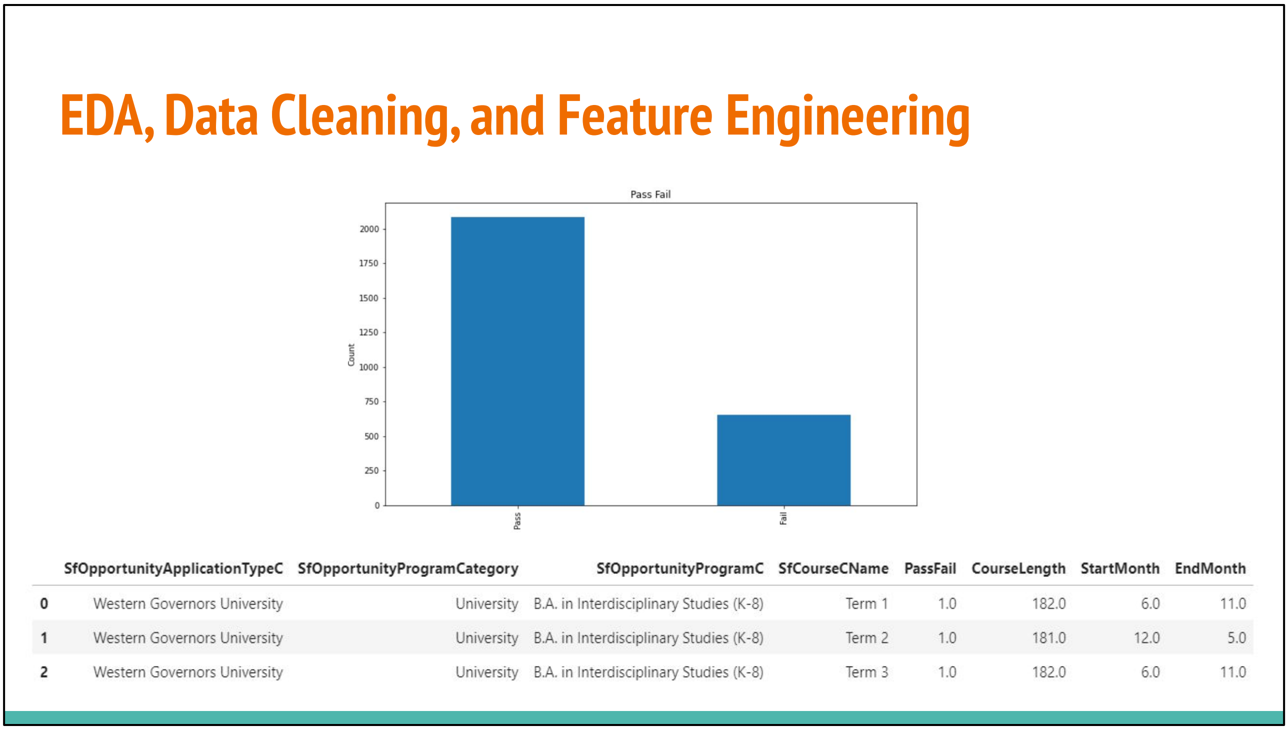


Modeling Course Outcomes

Michael Pajewski



EDA, Data Cleaning, and Feature Engineering

	SfOpportunityApplicationTypeC	SfOpportunityProgramCategory	SfOpportunityProgramC	SfCourseCName	PassFail	CourseLength	StartMonth	EndMonth
0	Western Governors University	University	B.A. in Interdisciplinary Studies (K-8)	Term 1	1.0	182.0	6.0	11.0
1	Western Governors University	University	B.A. in Interdisciplinary Studies (K-8)	Term 2	1.0	181.0	12.0	5.0
2	Western Governors University	University	B.A. in Interdisciplinary Studies (K-8)	Term 3	1.0	182.0	6.0	11.0

The After combining the three tables data set contained 7572 points of information on course but only 2736 data points had some form of grade.
So only those 2736 data points were used in the model.

There were a number of different grading options I determined

- Passing grades are A+ though C-, Passed, S, and SP
- All other options were considered failing
- I assumed blanks ment there was no information on the course grade
- From this I created a Binary predictor of 1 being passed the course 0 being failed the course based off these assumptions

One thing to note is the Data Set Imbalance you can see in the chart

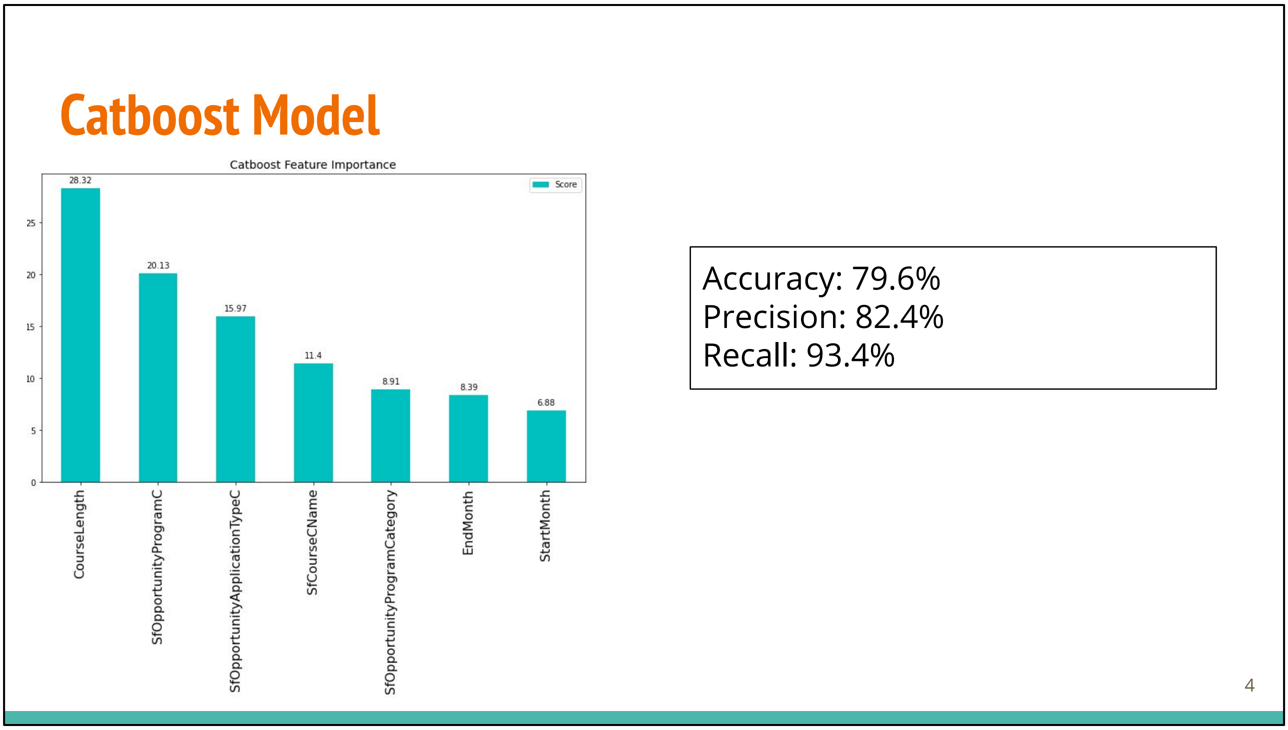
- 76% pass and 24% failure
- Somewhat of a data imbalance but not a server one

After some exploratory data analysis the following Features were chose for the model from the original data set

- SfOpportunityApplicationTypeC
- SfOpportunityProgramCategory
- SfOpportunityProgramC
- SfCourseCName

The following features were created to try and add meaning to original dataset features

- Course length in days
- Start month
- End month



Catboost Model

Accuracy: 79.6%
Precision: 82.4%
Recall: 93.4%

I choose to use Catboost because

- Highly categorical data set
- High cardinality of many data features
- Limited data set size
  - Catboost there is a special modification for small data sets to prevent overfitting
- Limited time to work on model trying to stay within the approximate 3 hour time limit
  - So Extensive feature engineering would be required for other model types

Model Parameters

- Working with a fairly small dataset and working with limited model tuning I stayed closed default parameters
- Used parameters of Iterations of 500, learning_rate of .1 and tree depth 2

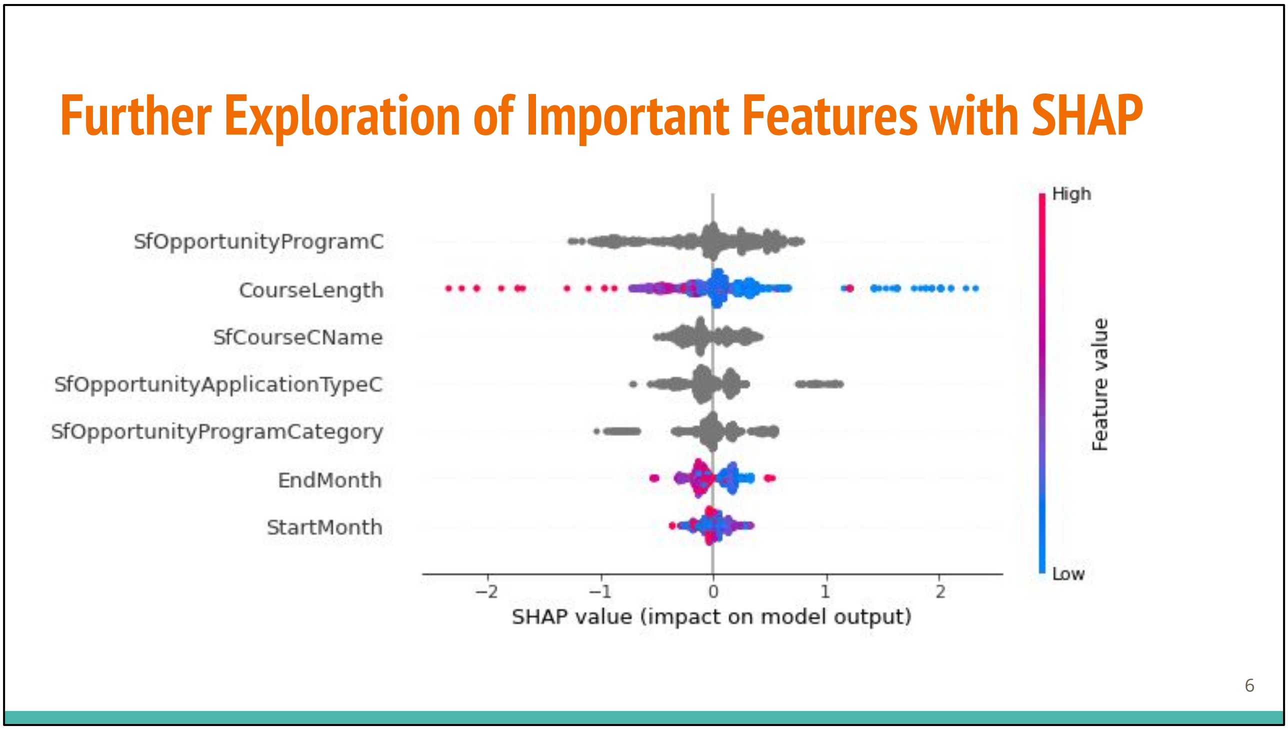
Model performance

- Accuracy: 79.6%
- Precision: 82.4%
- Recall: 93.4%

- instance where we consider that the cost of a false negative as high or when we consider predicting failure of a course when in reality the student would pass as very detrimental

Feature Importance

- We can see in the feature importance plot the 3 most important variables are
    - course length or number of days between start and end date
    - opportunity program
    - And Opportunity Application type



Further Exploration of Important Features with SHAP

Here is a more in depth feature importance plot shown through SHAP made with all data points in the training data
It is important to point out the SHAP values do not provide causality but helps to provide model explainability

- Feature importance: Variables are ranked in descending order slightly different order than the feature importance plot based on Prediction Values chang
- Impact: The horizontal location shows whether the effect of that value is associated with a higher or lower prediction.
- Original value: Color shows whether that variable is high (in red) or low (in blue) for that observation. Only for numeric features
- Correlation: A high level of the “Course length” content has a high and negative impact on the quality rating. The “high” comes from the red color, and the “negitive” impact is shown on the X-axis
  - End month also some positive weight is given to course that end in lower months ie the first few months of the year and a negative weight to course ending in the last few months of the year

Shap shows the positive and negative relationships of the predictors with the target pass fail