Predicting Student Outcomes

Client Problem

- Dropping out of high school carries a high cost to students
- General ‘stay in school’ messaging is likely less effective than more targeted interventions
- Schools and districts are under pressure to increase graduation rates
- High dropout rates can hurt school climate and teacher morale

The Data Science Problem

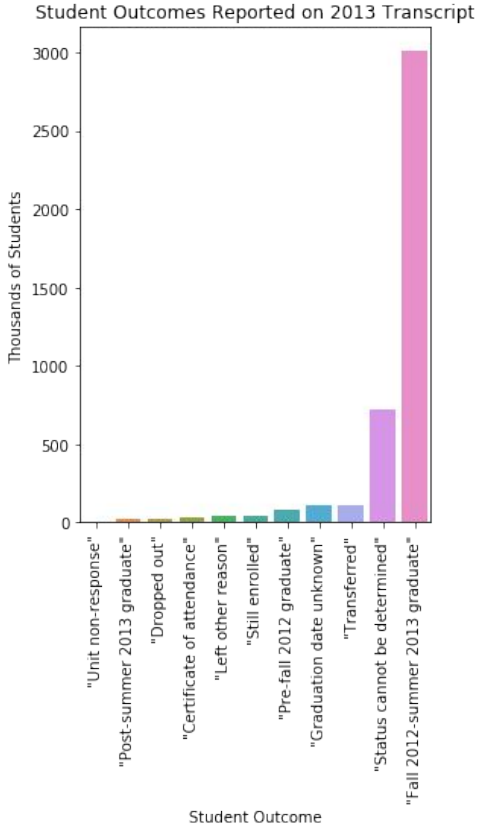
- We would like to classify students into ‘dropout’ and ‘non-dropout’ categories
- We would like to assign a probability of dropout to each student.

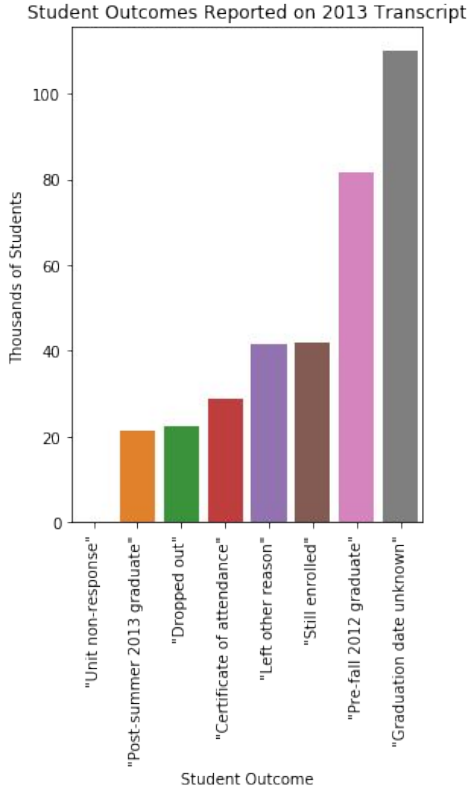
Data Wrangling

- Survey Data from the National Center for Education Statistics (NCES)
- Longitudinal study of 2009-2016
  - Followed about 23,000 ninth graders starting in 2009 over seven years
  - Complex sample design was used to make sure that the survey population was represented of the US as a whole.
  - A version with certain data labeled ‘suppressed’ is freely available without permission
- Data contained in several formats (like R and STATA), including CSV file
- In addition to importing survey answers, also import weight columns
- Also import documentation text which connects the numbers present in the data tables with the corresponding answers to survey questions

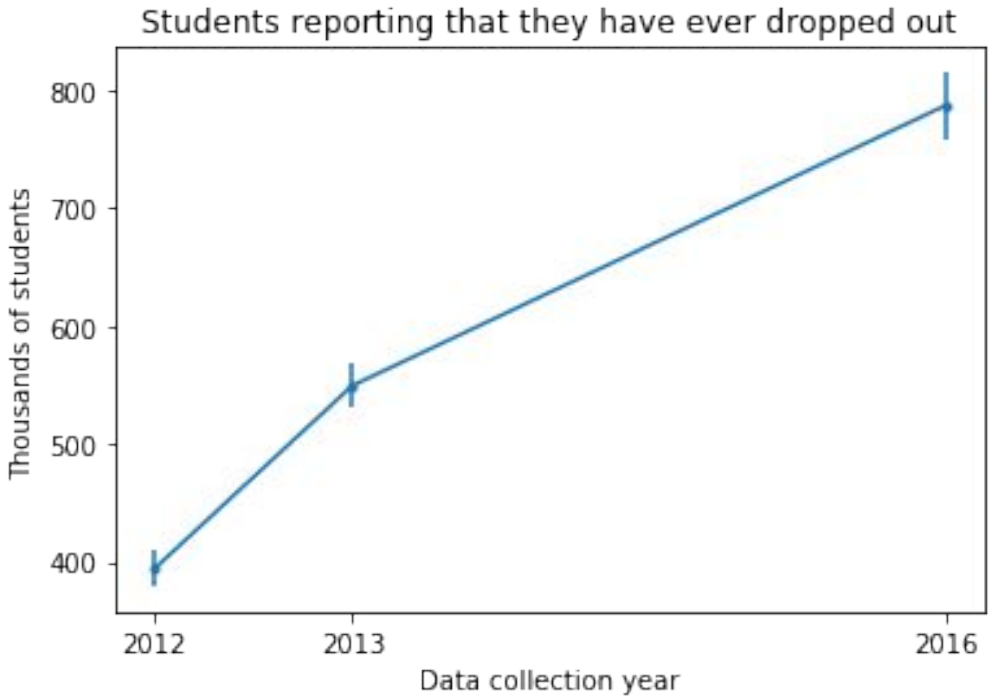
Exploratory Analysis

- Outcomes as listed on student transcripts gives very rosy outlook.
- Total students that end with ‘dropout’ on their transcript is 20 thousand out of about 4.2 million freshman.
- Large number of ‘unknown status’ students





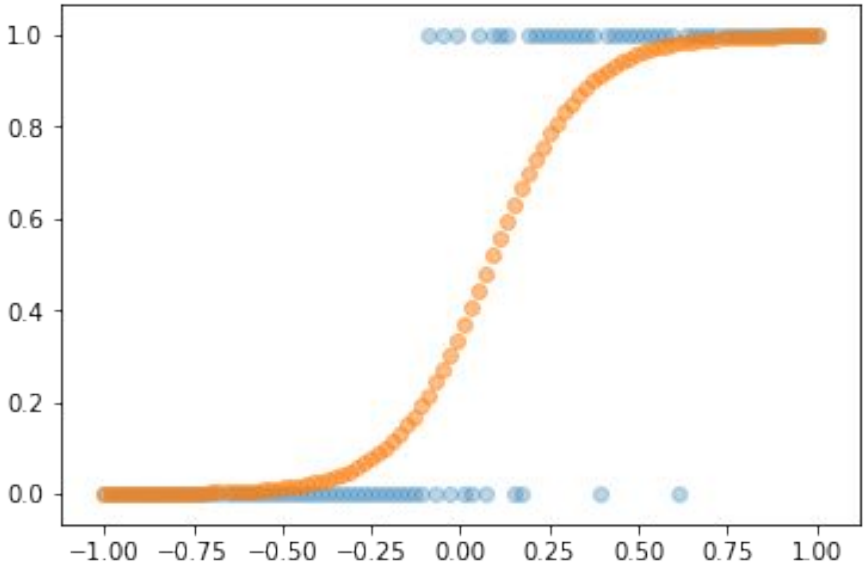
Exploratory Analysis



- When we look at dropout as reported by students and parents, we see that a much greater number of students have dropped out
- Starting with the original cohort's junior year, we see that more students dropped out of high school each year.
- We use the 2013 'have you ever dropped out of high school' question as labels for dropout/non-dropout

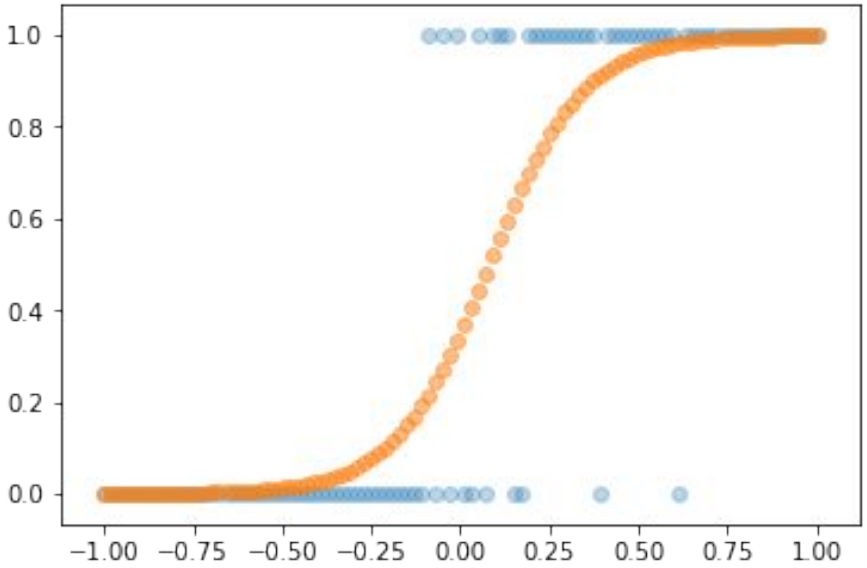
The Machine Learning Model - Logistic Regression

- Logistic regression is a highly interpretable model, useful in categorization.
- The model fits the given data using a logit function that takes on values ‘0’ and ‘1’, as well as values in between.
- The coefficients created by the model are related to the change in the odds that the correct class is ‘1’ for a unit change in the dependent variable.

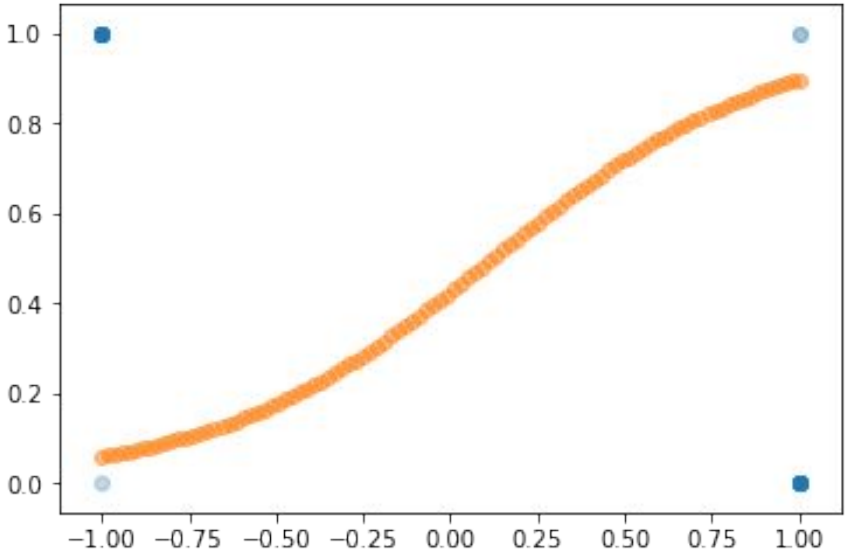


Logistic Regression - Continuous Data

- Logistic regression can be used to model data that takes on two values.
- When the independent variable is continuous, a smooth curve can be drawn.
- The curve represents the probability that the correct class is ‘1’.

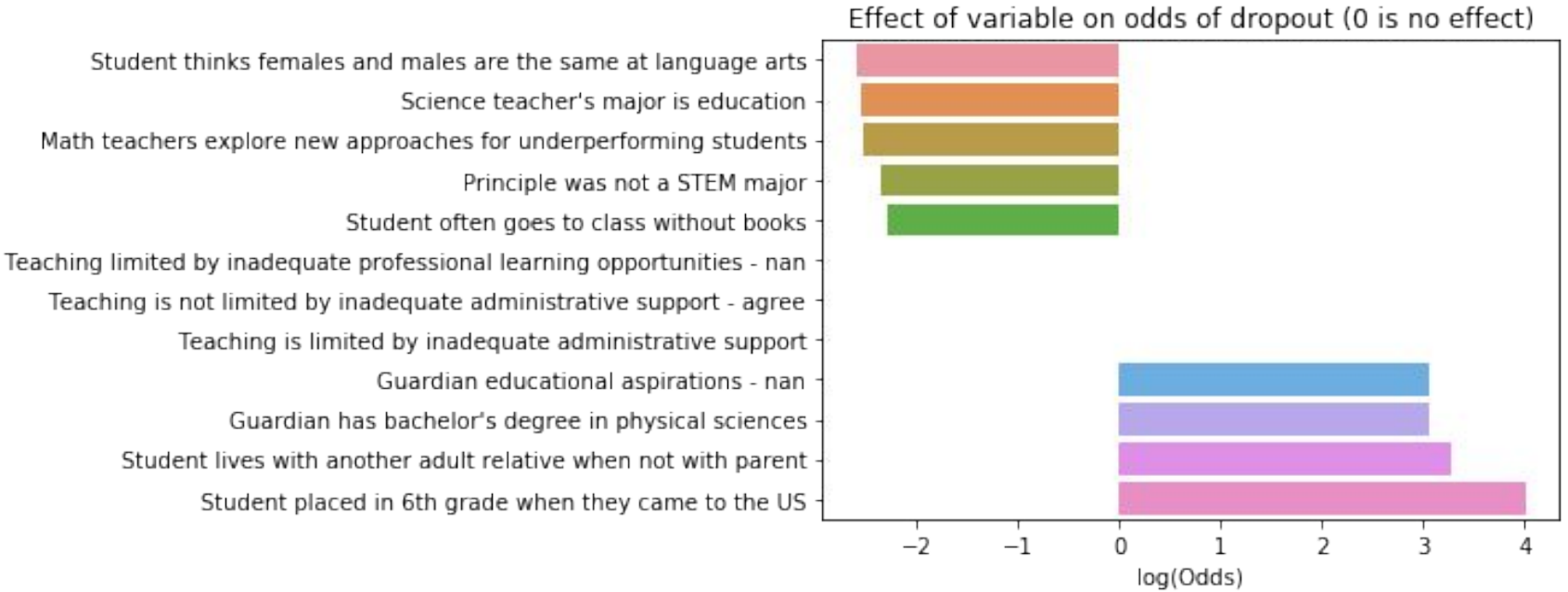


Logistic Regression - Discrete Data



- Logistic regression can also model discrete data.
- The model predicts the probability that each value of the independent variable is ‘1’.
- The predicted probabilities for the independent variable no longer have meaning.

Results - Top positive and negative coefficients as well as three neutral coefficients.



Results - Precision and Recall for different models

Model	Precision	Recall	F1
Baseline	0.32	0.40	0.36
Balanced Class Weights	0.28	0.47	0.48
Random Undersampling	0.20	0.62	0.30
Random Oversampling	0.30	0.45	0.36
F1 Score Optimization	0.56	0.29	0.38
Balanced Class weights and F1 Score Optimization	0.27	0.69	0.39

Client Recommendations - Example: Recall 80%, Precision 25%

- A school of 725 students will have approximately 100 potential dropouts
- 80% recall means that 80 out of the 100 dropouts are identified
- 25% precision means that 320 students are identified as potential dropouts
- This is nearly 45% rather than 15% of the student population
- This would be a good route for a school if the intervention were aimed at entire grades

Client Recommendations - Example: 50% Recall, 40% Precision

- 50% recall means that 50 out of 100 dropouts are identified
- 40% precision means that 125 students are identified as potential dropouts
- These are more manageable numbers for a targeted intervention meant to treat students in small groups
- An even more precise model would focus on a handful of students

Client Recommendations - Best of both worlds

- We strongly recommend including humans as part of the workflow
- Combining human scanning, and knowledge of students as individuals, achieves the best of both worlds
- We can attempt a high recall model to narrow down the list of possible dropouts
- Use teachers, counselors, and administrators familiar with the student to increase the precision

Future Improvements

- Restricted Use Files
  - More detailed data
  - School level data as well as student level
  - Partner with schools and districts
- Different Approach
  - Anomaly detection
  - Feature selection and interaction