High School Graduation Rates

Introduction

- Data set consists of info on every county in the U.S.
- Contains features such as race, homeownership, health, death rates, income inequality, more (~ 100 + unique features)
- Target feature: Binned High school graduation rates

Modeling Objectives

- Model high school graduation rates and understand factors that are likely to be predictive of graduation rates.
- Initally attempted to predict high school graduation rates as multiclass classifier problem, analysis follows.
- The resulting model performed better than chance, to improve accuracy the analysis was repeated binning graduation rates into counties greater than and less than 90% (binary classifer).

Looking at Correlations

Top 3 Positive Correlations:

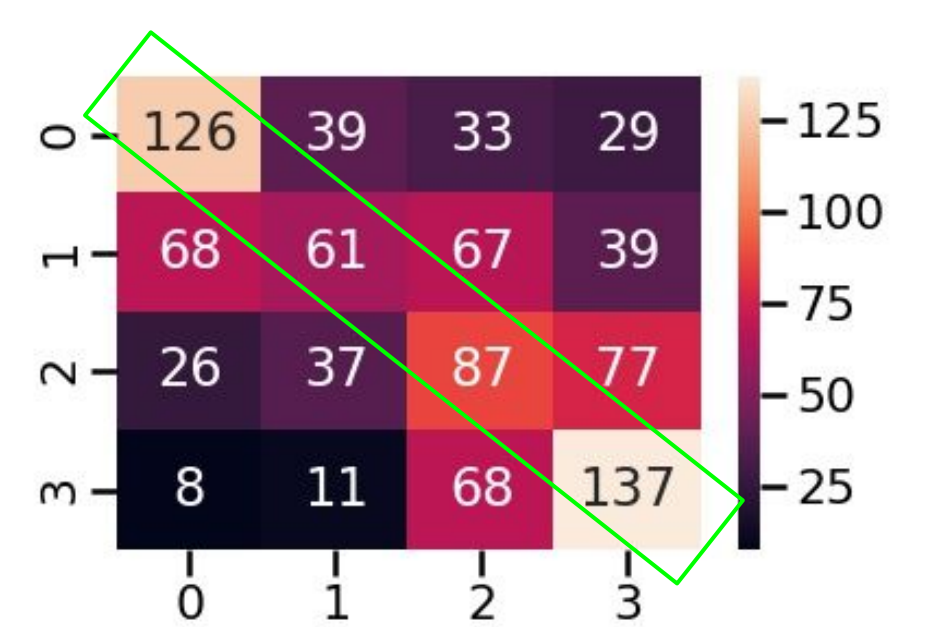
1. Percent non-hispanic white (Correlation: 0.31)
2. Home ownership (Correlation: 0.27)
3. Percentage rural (Correlation: 0.22)

Top 3 Negative Correlations:

1. Children in single-parent households (-0.32)
2. Severe housing problems (-0.31)
3. Percentage of households with high housing costs (-0.31)

Multiclass modeling

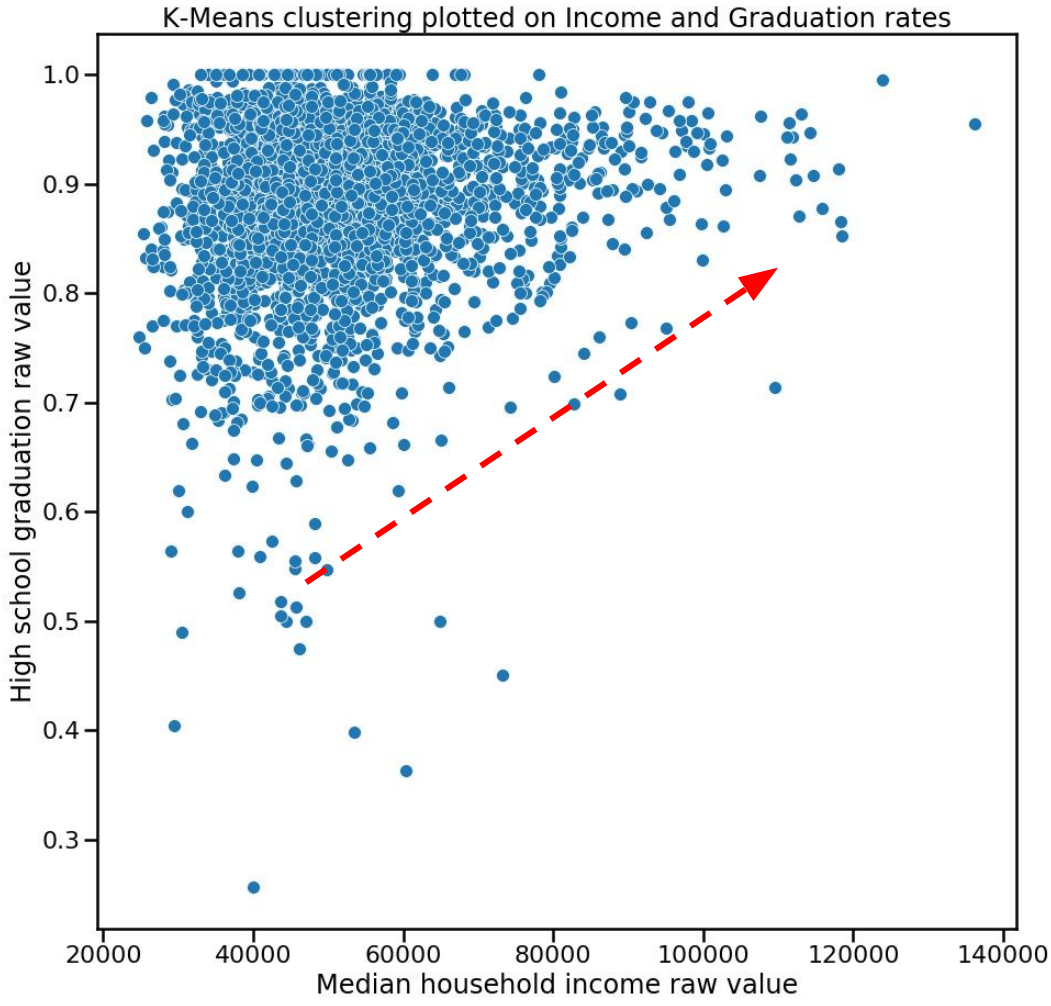
- Graduation rate split by quantiles
  - 0-84%
  - 84-89%
  - 89-93%
  - 93-100%
- Best model
  - KNN
- Performance
  - accuracy = 45%
  - higher than random chance (25%)
  - Better at extreme highs/lows



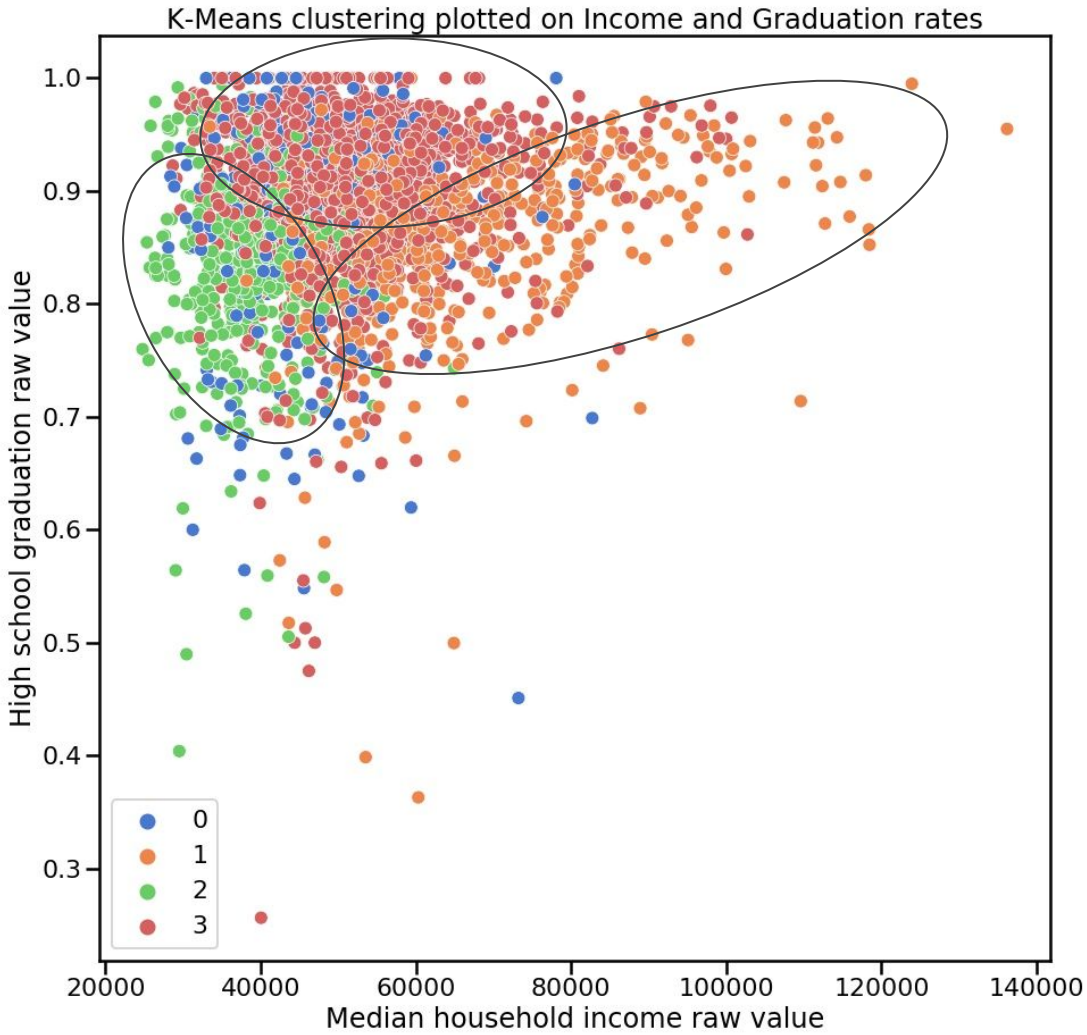
Relationship between income and graduation rates

High income
- High grad rate
- Less variation

Low income
- Lower grad rate
- Higher variation



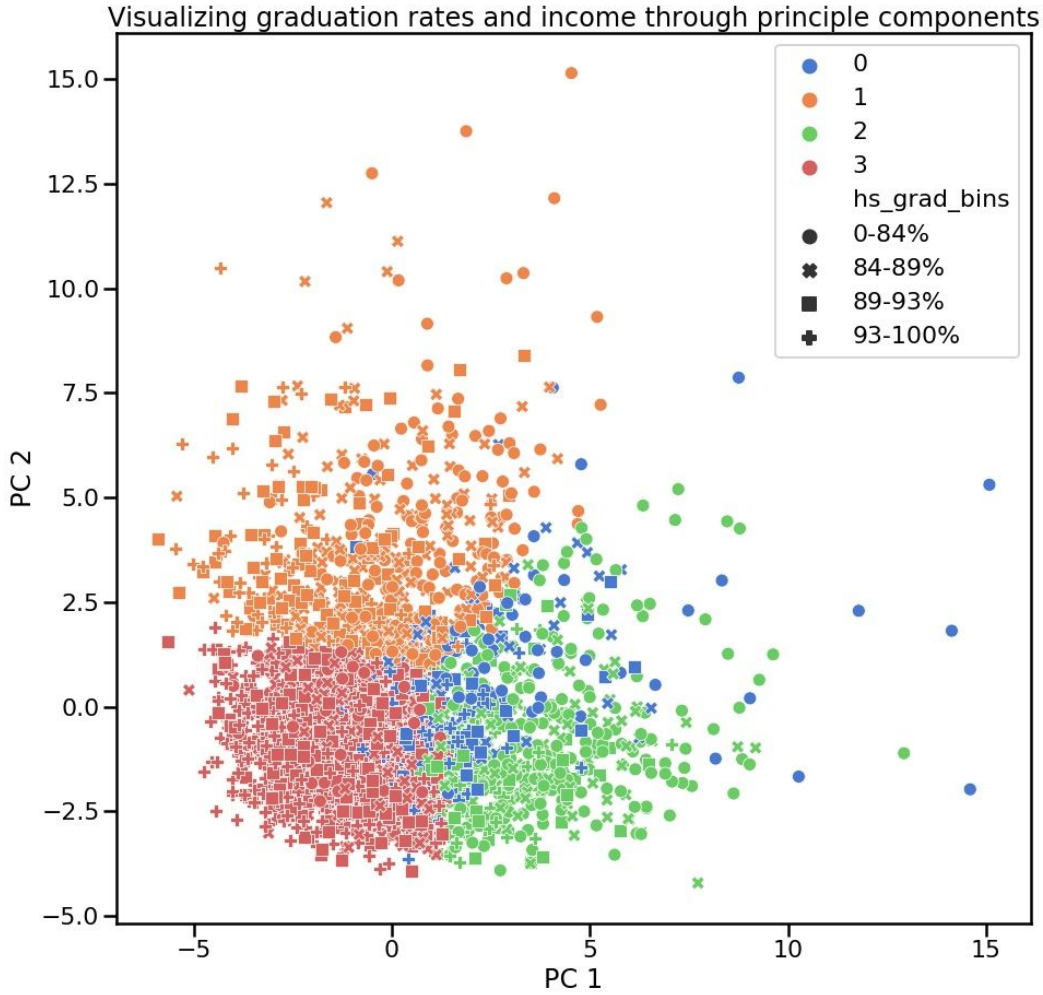
K-means clustering overlay



K-Means with PCA

K-means seems to map with high and low grad rates reasonably well at K=4

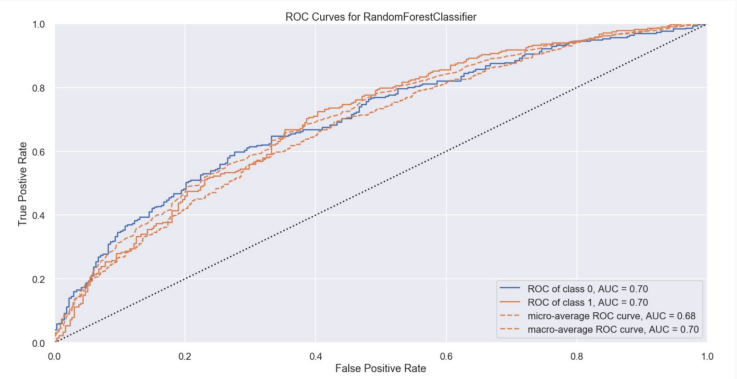
This might be improved by reducing graduation rate bins to 3 and K to 3

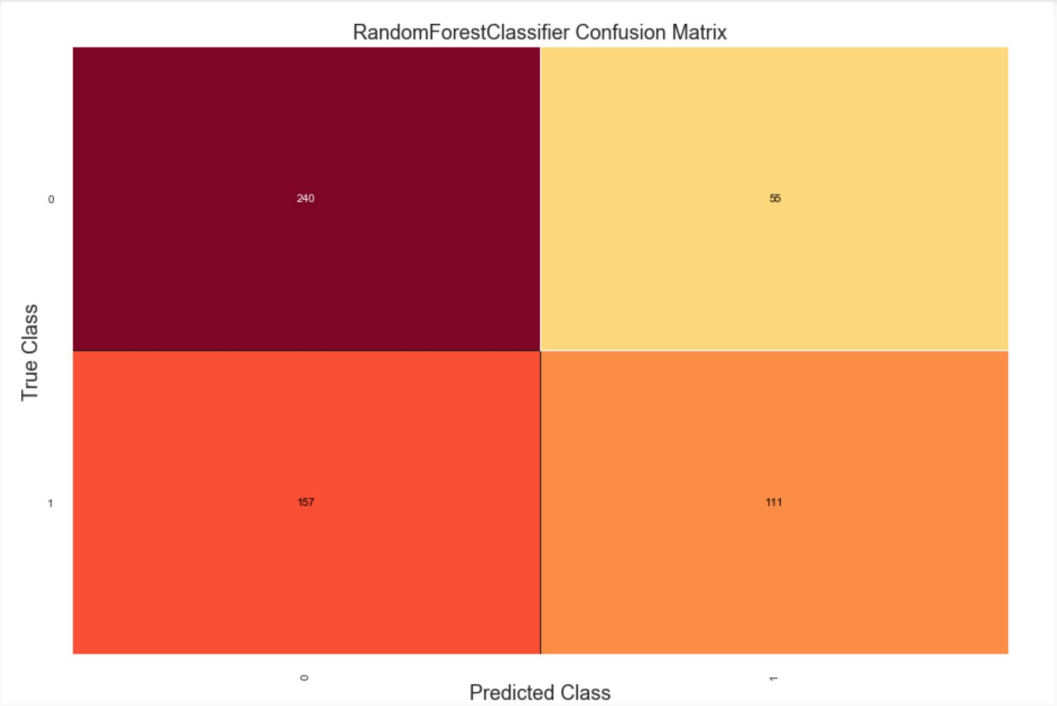


Model performance as a binary classifier problem

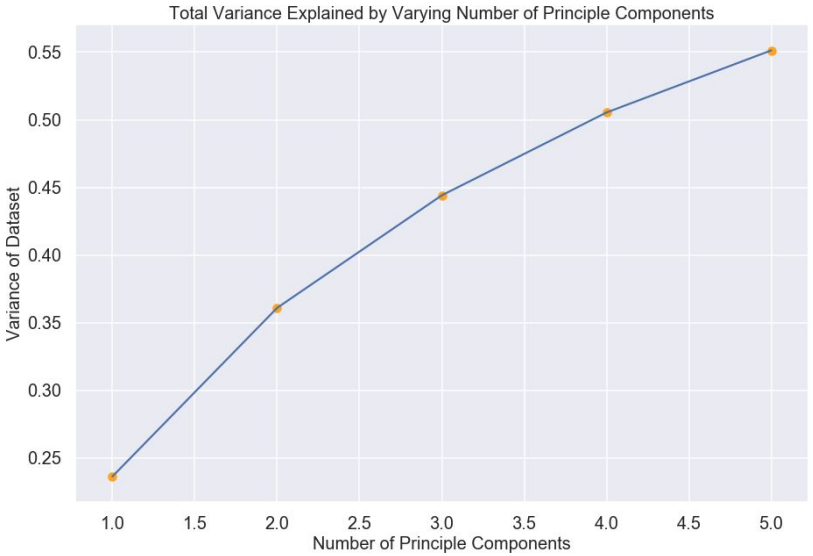
Random Forest

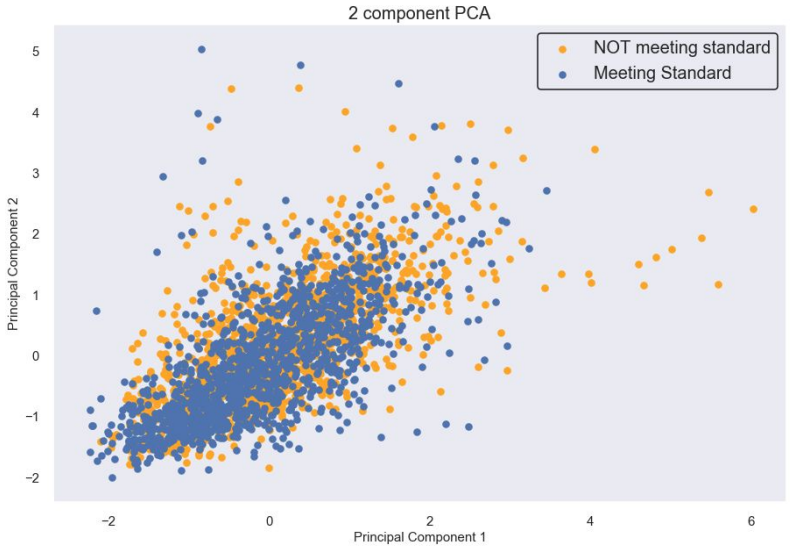
	precision	recall	f1-score	support
0	0.76	0.76	0.76	295
1	0.74	0.73	0.73	268
accuracy			0.75	563
macro avg	0.75	0.75	0.75	563
weighted avg	0.75	0.75	0.75	563





Principal Component Analysis on Raw Data

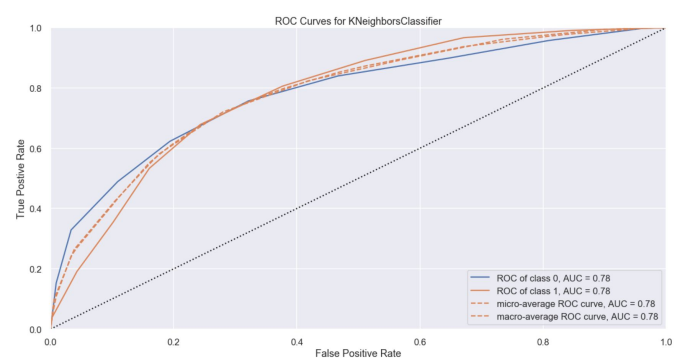


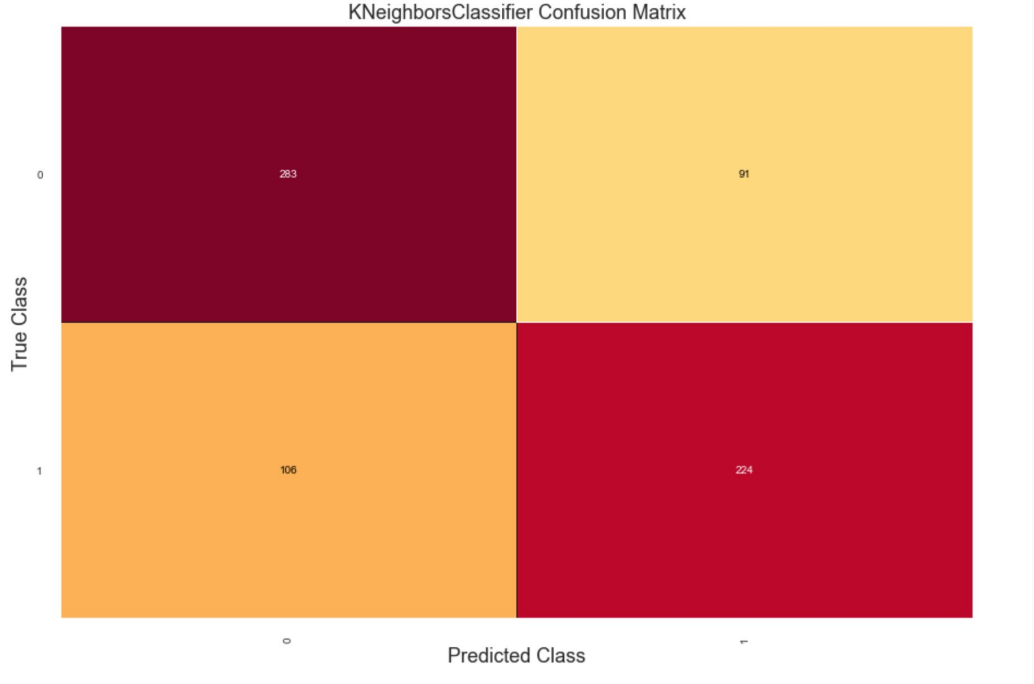


KNN Model

Accuracy:0.7116477272727273
F1: 0.687211093990755

	precision	recall	f1-score	support
0	0.72	0.74	0.73	374
1	0.70	0.68	0.69	330
accuracy			0.71	704
macro avg	0.71	0.71	0.71	704
weighted avg	0.71	0.71	0.71	704

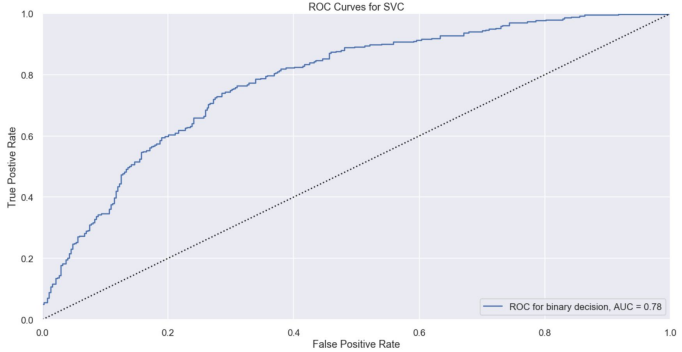


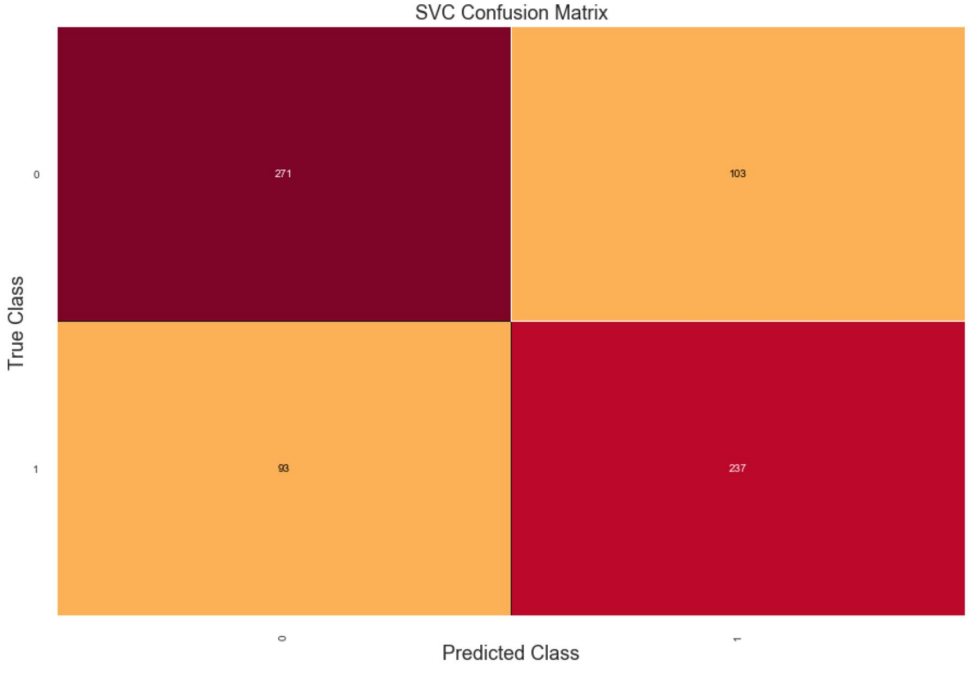


SVM Model

Accuracy:0.7215909090909091
F1: 0.7074626865671643

	precision	recall	f1-score	support
0	0.74	0.72	0.73	374
1	0.70	0.72	0.71	330
accuracy			0.72	704
macro avg	0.72	0.72	0.72	704
weighted avg	0.72	0.72	0.72	704





Conclusions

1. Graduation rates correlate positively rural areas, home ownership, non-hispanic white populations, negatively with high housing costs, and single parent households.
2. KNN is best predictor in multiclass scenario, Random Forest performed best in binary scenario
3. KMeans clustering appears to show good grouping at extremes but there is overlap between two middle groups when comparing to graduation rates.
4. Clustering and multiclass modeling may be improved by reducing graduation bins to 3