Capstone 1 Final Report
Springboard Curriculum Section 10.6

UNDERSTANDING LOW COMPLETION RATES AT COMMUNITY COLLEGES

William C. Webb

INTRODUCTION

Community Colleges in the US play the vital role of bringing higher education and vocational training to a broad spectrum of the general population. Unfortunately, community colleges suffer extremely low completion rates. In the US, only 23% of first-time degree/certificate-seeking students at two-year public colleges complete their programs within six years, compared to 47% at four-year public colleges (NCES 2019).

Low completion rates at community colleges disproportionately impact economically-disadvantaged students. Community colleges function as the only economically-viable conduit to higher education for many low-income students. Almost half of all low-income Americans first enroll in a community college to pursue higher education (Mann-Levesque 2019). Individuals without any type of degree or certificate beyond a high school diploma face substantially reduced earning potentials (Belfield and Bailey 2017, Mann-Levesque 2019). Without access to higher education or vocational schools, low completion rates at community colleges reduce economic mobility.

Through their vocational programs, community college help train a skilled workforce. The projected growth rate for 43 occupations requiring an Associate's Degree or certificate for entry-level positions is expected to increase, according to the US Bureau of Labor Statistics (BLS 2019). However, low completion rates at community colleges may result in skilled positions remaining unfilled and therefore weaken local economies.

Improving community college completion rates is vital not only for individual economic mobility but also maintaining a productive and skilled workforce. In this project, I identified important predictors of completion rate at community colleges. I hypothesized that the level of student support, such as the availability of financial aid and the quality of student services, represent the strongest predictors of completion rate.

METHODS

Data Wrangling

The College Scorecard (U.S. Department of Education 2019) contains a collection of online tools and data for comparing title IV institutions, including community colleges offering undergraduate degrees. Title IV institutions support US federal student financial aid programs for students. In addition to completion rates, the data describe each institution in numerous ways including incoming student SAT scores, percentage of degrees in various fields of study, cost of study, faculty salaries and many other metrics. The College Scorecard posts data reported annually between 1996 and 2017 for download in .csv format and related documentation https://collegescorecard.ed.gov/data/documentation/ . Collectively, these 22 files contain 2.46 GB of data, averaging 111.18 MB each.

I used Jupyter notebooks and Python (v. 3.7.4.) with associated libraries to conduct data wrangling and all subsequent analyses. As a first step, I read the original .csv files into a Jupyter notebook and combined them into a single DataFrame containing 154,228 cases (rows) and 1977 features (columns).

Data Cleaning

As part of the initial data cleaning effort, I selected cases for inclusion in the analysis and performed feature selection. I began this data cleaning process by manually excluding features and cases from inclusion in the analysis. I excluded features and cases primarily based on the research question and personal domain knowledge gained as a part-time community college instructor.

My initial goal for case selection was retaining cases pertaining to public community colleges, where an Associate's Degree was the highest degree awarded. In the original full dataset, the number of cases where the highest degree awarded was the Associate's Degree represented 24% (37108 cases out of 154,228 original cases) (Fig. 1).

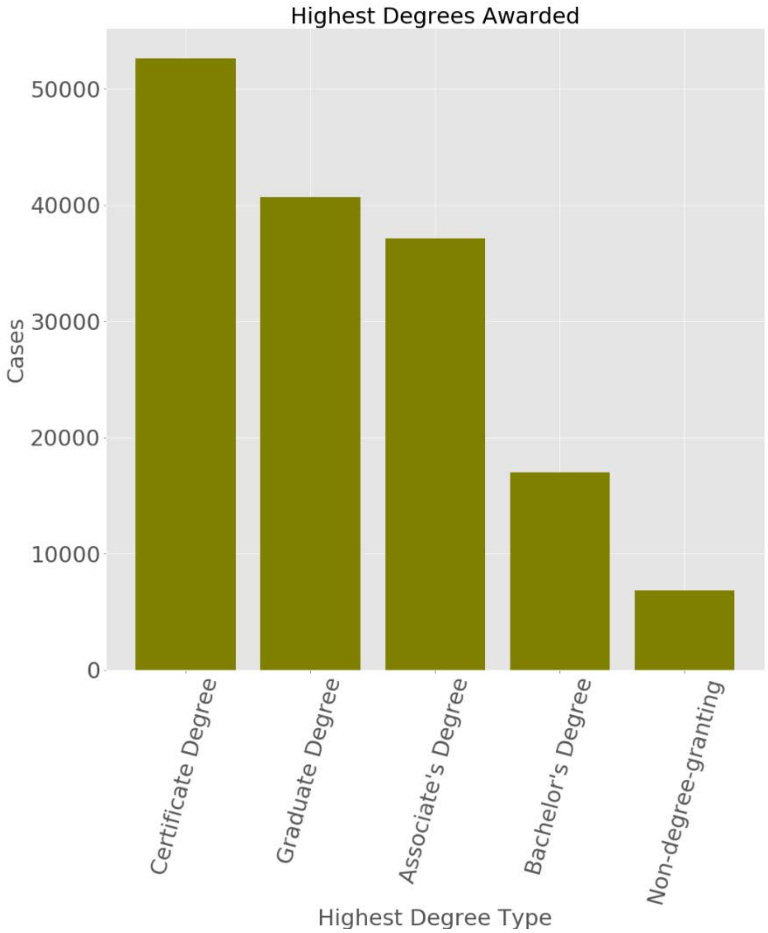


Fig. 1. Comparison of the number of cases of colleges by highest degree offered in the College Scorecard data, between the academic years of 1996-97 through 2017 – 2018.

I filtered the original full data to select only public colleges, reducing the original 154,228 cases to 46,177 cases pertaining to public colleges and 1977 features. I then filtered the cases pertaining to public colleges for those where the highest degree awarded was the Associate's Degrees, representing 22655 cases and 1977 features, amounting to 15% of the original 154228 cases. Of the 22,655 cases pertaining to public colleges where the highest degree offered was the Associate's Degree, 24% (5515 cases) predominantly offered Certificate Degrees but also offered some Associate's Degrees.

I included features that I deemed most relevant to the research question resulting in a working DataFrame containing 231 features and 22,655 cases. For the target/dependent variable/predictor (hereafter 'target'), I chose the most-relevant completion measure, abbreviated as "C150_L4". The target was defined by the College Scorecard as: "Completion rate for first-time, full-time students at less-than-four-year institutions (150% of expected time to completion)", and referred to hereafter as "completion rate". With the exception of the target variable, I excluded any additional features related to measures of completion. I also excluded features related to student debt, loan repayment rates, sex ratios, post-enrollment earnings, and other features I perceived as unrelated to the research question. I added the year that each case was reported as an additional feature to the DataFrame as well.

I followed general guidelines for addressing missing values. For example, I removed all features composed of > 25% NaN, which resulted in a working DataFrame containing 86 features and 22,655 cases. I replaced all remaining missing values (NaN and zeros) in the 86 features with the mean value for each feature, respectively. I identified and removed two cases where the value for the categorical feature 'Open Admissions Policy ('OPENADMP') corresponded to "Does not enroll first-time students", resulting in a working DataFrame with 86 features and 22653 cases. Because they could not be imputed, I removed an additional 1076 cases where the categorical feature 'OPENADMP' contained NaN values, resulting in a working DataFrame containing 86 features and 21577 cases.

Linear Regression and Model Assumptions

Prior to modeling the relationship between completion rate and the explanatory features, I maximized the data's conformity to meet the assumptions of linear regression (Appendix) to ensure that inferences from linear regression would be valid. These efforts included standardizing (sklearn.preprocessing.StandardScaler) and log-transforming the data (np.log), as well as removing 387 outliers (sklearn.ensemble.IsolationForest). Outlier removal resulted in a working DataFrame with 21,190 cases/rows and 86 features. I also removed 1387 cases containing repeated values (0.05) for the target 'C150_L4' which I believed erroneous, resulting in working DataFrame with 19803 vases and 86 features. After a preliminary linear regression analysis (statsmodels.regression.linear_model) produced multicollinearity warnings, I conducted feature selection using variance inflation factor (VIF) scores (statsmodels.stats.outliers_influence.variance_inflation_factor) and removed continuous features with VIF scores > 5. This resulted in removal of 34 features and a working DataFrame containing 19803 cases and 52 features including the target - completion rate. I used this DataFrame for a linear regression analysis relating completion rate with other features, and the results of this analysis for statistical inference.

Machine Learning

Subsequent to the traditional linear regression analysis used for statistical inference, I compared the performance of seven linear regression machine learning algorithms in their abilities to generalize to unseen data. I initially used linear regression (sklearn.linear_model.LinearRegression) with ten-fold cross validation (sklearn.model_selection.cross_val_score). I tuned hyperparameters using nested cross validation (sklearn.model_selection.GridSearchCV, sklearn.model_selection.cross_val_score) for the remaining six algorithms. I then used regularizing algorithms (Ridge, Lasso and ElasticNet; sklearn.linear_model) to minimize the potential undue influence of any large regression coefficients. Finally, I generated models using the regression algorithms for random forests (sklearn.ensemble.RandomForestRegressor), nearest neighbors (sklearn.neighbors.KNeighborsRegressor) and support vector machines (sklearn.svm.SVR).

RESULTS

The overall mean completion rate for cases representing public colleges offering the Associate's Degree as the highest degree awarded was 0.24 (± 0.14 SD) (Fig. 2).

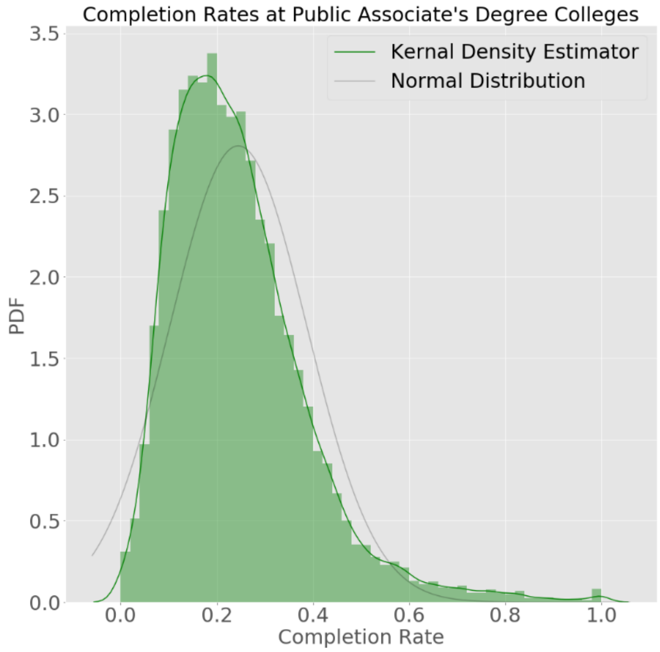


Fig. 2. Distribution of completion rate for cases representing public colleges offering the Associate's Degree as the highest degree awarded.

Linear Regression

After correcting for potential multicollinearity, linear regression using the remaining features explained less than a third of the variation in completion rate (R^2_{adj} = 0.29, DF = 51, P <0.01).

Parameter Estimates

I used model results from linear regression to infer the most important parameters related to completion rate (Appendix Table 1). For statistical inference, I adopted the convention of alpha = 0.05 for the type I error rate unless stated otherwise. The percentage of degrees awarded in agriculture (PCIP01) had the strongest positive relationship (ß = 0.16) with completion rate (Fig. 3, Fig. 4) while the share of part-time degree-seeking students (PPTUG_EF) had the strongest negative relationship (ß =- 0.27) with completion rate (Fig. 3, Fig. 5). Other parameters with strong positive relationships with completion rate included the percentage of degrees awarded in construction trades (PCIP46; ß = 0.13; Fig. 3), the percentage of degrees awarded in precision production (PCIP48; ß = 0.10; Fig. 3) and the indicator for main campus (MAIN; ß = 0.10; Fig. 3). Additional parameters with strong negative relationships with completion rate included awarding predominantly Associate's Degrees compared to certificates (PREDDEG; ß = -0.19; Fig. 3), the percentage of students whose parents' highest education was high school (PAR_ED_PCT_HS; ß = -0.11; Fig. 3) and the percentage of degree awarded in education (PCIP13; ß = -0.10; Fig. 3).

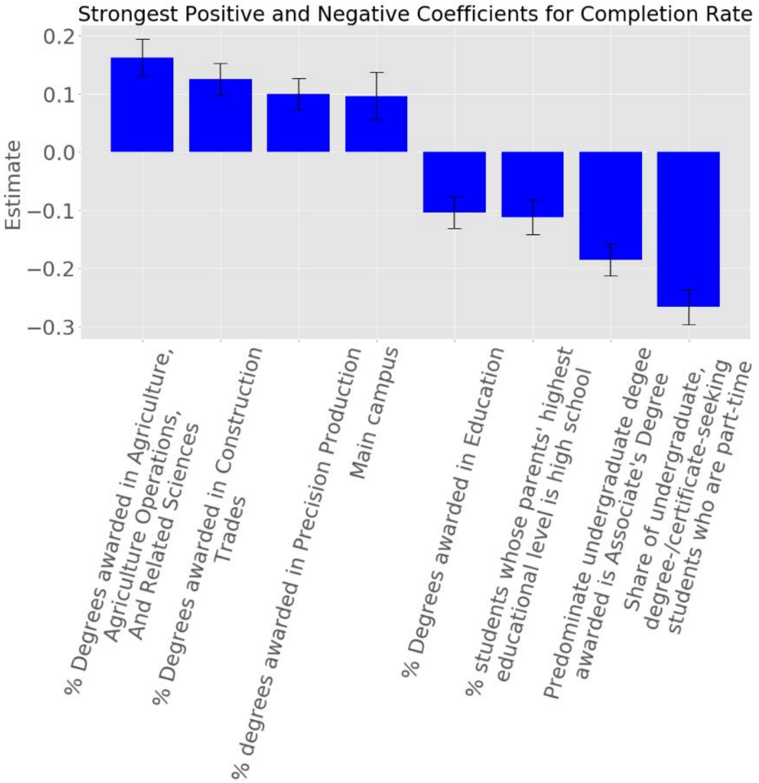


Fig. 3. Strongest positive and negative linear regression coefficients.

Machine Learning

The best-performing machine learning algorithm, as measured by the mean coefficient of determination for prediction was random forest regression (R^2 = 0.20 ± 0.13 SD) and the lowest-performer was support vector machine regression (R^2 = 0.08 ± 0.14 SD) (Fig. 6).

DISCUSSION

Both statistical models and machine learning models employing linear regression identified clear relationships between completion rate and other features in the College Scorecard data. Although none of the models explained more than 30% of the variance in completion rate, the results provide new insights regarding the relative importance of individual features for modeling completion rate using College Scorecard data.

Interpreting the Strongest Positive Predictors

Contrary to the research hypothesis, the strongest positive predictors of completion rate were specific types of individual degree programs, rather than parameters associated with student support. In particular, the percentage of degrees awarded in skilled-trade degree programs including agriculture, construction and precision production represented the three strongest positive predictors of completion rate. Compared to traditional academic programs of study, skilled trade programs may promote higher completion rates by enhancing student motivation with more concrete graduation outcomes, reduced ambiguity regarding student achievement due to the practical nature of the subject material, a greater tendency for highly-motivated students to self-select prior to enrollment, and potentially fewer course requirements. In addition, the nature of learning skilled trades often requires smaller class sizes, close mentorship, and student cohorts which research shows all promote higher student success.

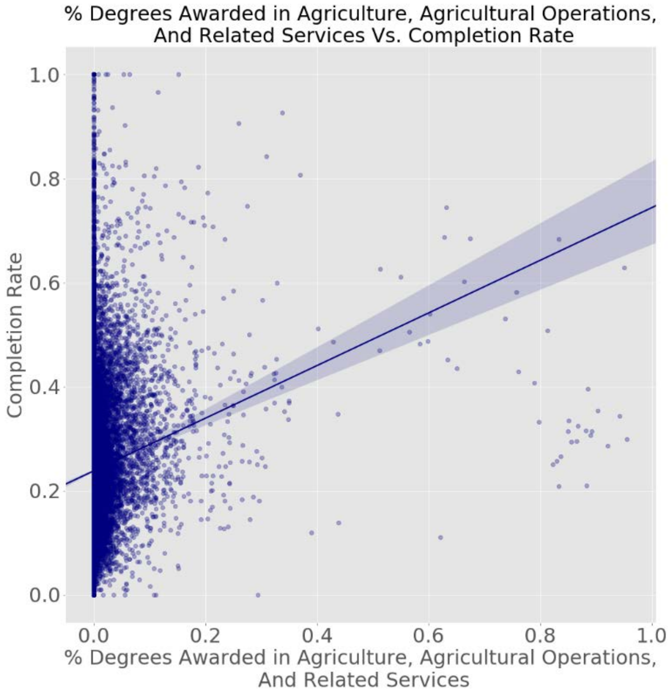


Fig. 4. Relationship between completion rate and the strongest positive predictor: percentage of degrees awarded in agriculture, agricultural operations and related services (PCIP01).

Although not in the top three positive predictors, the fourth-ranked positive predictor- the main campus indicator – does provide some support for the research hypothesis that investment in student support increases completion rate. Main campuses typically provide greater access to student support services compared to newer, smaller and less well-established satellite campuses. Instructional expenditures per full-time equivalent student (INEXPTFTE') ranks as the fifth highest-ranking positive predictor and supports the research hypothesis.

Interpreting the Strongest Negative Predictors

The top negative predictors of completion rate shed light on the challenges faced by community college students and provide potential guidance for administrators on addressing those needs. In terms of absolute value, the percentage of part-time degree-seeking students was the strongest predictor of completion rate. The negative influence on completion rate of this factor is likely due to the longer matriculation time required for part-time degree-seeking students and may reflect that these students likely face greater external barriers such as work or family commitments that degrade their ability to focus on education. The second strongest negative predictor of completion rate was the campus indicator for the predominate awarding of Associate's Degrees. Not surprisingly, this result suggests that satisfying the requirements for an Associate's Degree present a greater hurdle than completing a certificate which often requires fewer courses to complete.

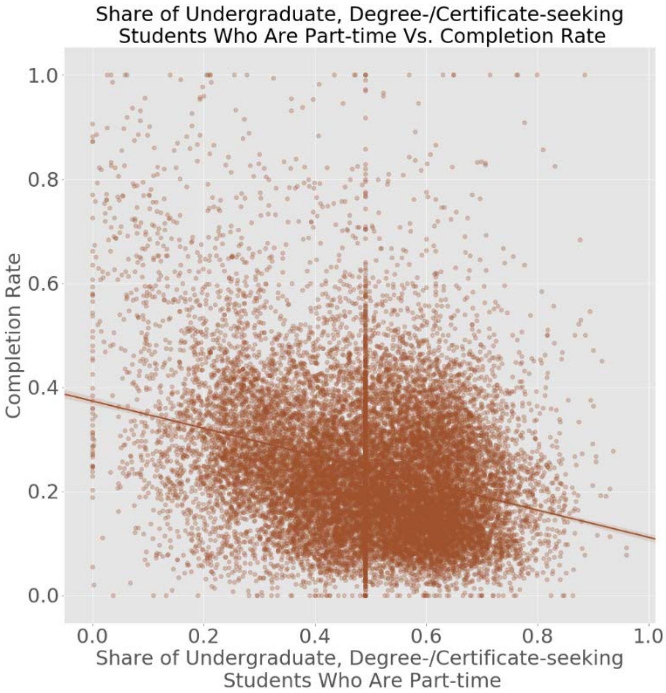


Fig. 5. Relationship between completion rate and the strongest negative predictor: share of undergraduate, degree-/certificate-seeking students who are part-time (PPTUG_EF).

The third and fourth strongest negative predictors of completion rate confirm previous research findings and provide some ambiguity, if not irony. Educational research shows that students often face substantial challenges achieving higher levels of education than their parents. The third strongest negative predictor of completion rate, the percentage of students whose parents' highest educational level was high school – aligns with those research findings. The fourth strongest negative predictor of completion rate – the percentage of degrees awarded in education – presents some ambiguity and irony. Clear explanation for this result seems evasive. Perhaps degrees in education maintain high graduation standards whose aim is to produce high-quality educators.

Failure to adequately meet all the assumptions of linear regression (see Appendix) suggests that additional transformations or different models might improve model fit and overall performance. For example, distribution fitting could reveal patterns that suggest better-fitting, non-linear models or suggest feature transformations that enable a better match to a linear model. Non-parametric models also present potential alternatives for improving model fit.

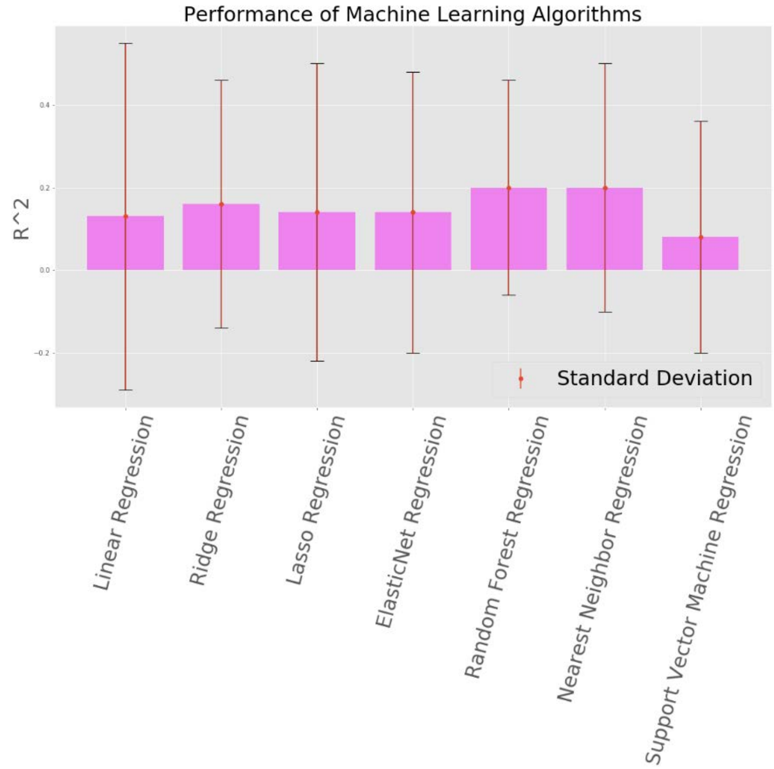


Fig. 6. Performance of machine learning algorithms for predicting completion rate.

REFERENCES

Belfield, C., and T. Bailey. 2017. The Labor Market Returns to Sub-Baccalaureate College: A Review. Center for Analysis of Postsecondary Education and Employment, Teachers College, Columbia University, New York, NY.

BLS. 2019. Occupational outlook handbook. United States Bureau of Labor Statistics.

Mann-Levesque, E. 2019. Improving community college completion rates by addressing structural and motivational barriers. Brown Center on Education Policy.

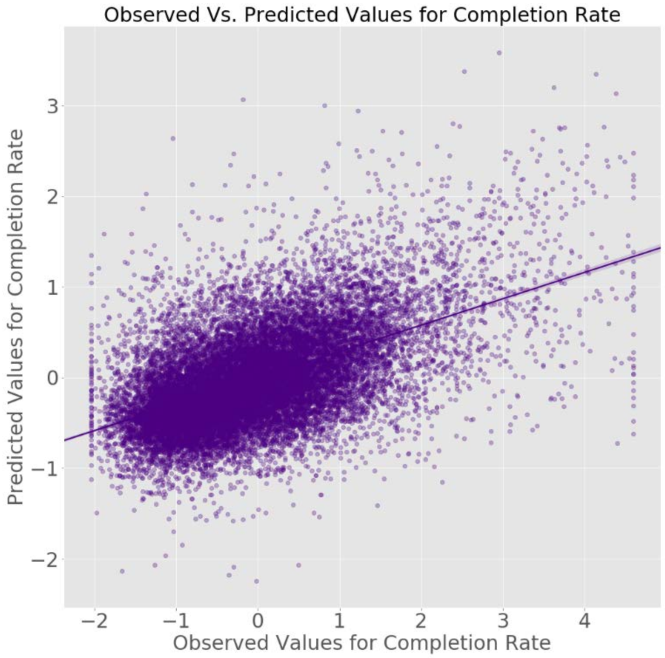
NCES. 2019. Digest of education statistics. National Center for Education Statistics.

U.S. Department of Education. 2019. College Scorecard Data.

APPENDIX

Assumptions of Linear Regression

Descriptions of the assumptions of linear regression differ somewhat between sources, but in general include the following: 1) a linear relationship between the features and the target; 2) the errors/residuals are independent of each other; 3) homoskedasticity/constant variance of the errors/residuals; 4) normally-distributed errors/residuals including a mean of zero; 5) absence of substantial multicollinearity between features.



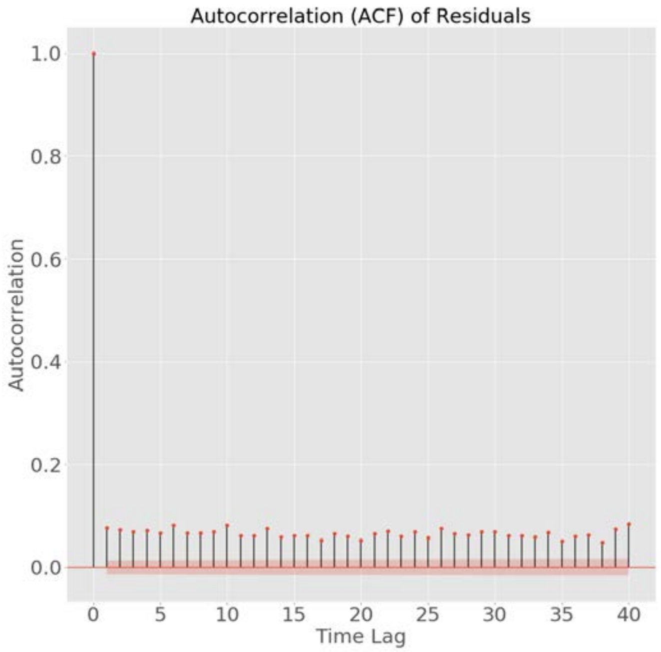
Appendix Fig. 1. Observed values of completion rate plotted against predicted values of completion rate generated from a linear regression model. Data are log-transformed and standardized. The overall pattern suggests the data meet the linearity assumption of linear regression.

Data Cleaning for Linear Regression Assumptions

To minimize potential violations of the normality assumption, I standardized (sklearn.preprocessing .StandardScaler) and log-transformed the data (np.log). I also identified and removed outliers since their presence sometimes contributes to violations of normality

Preliminary Linear Regression Results

I conducted a preliminary analysis modeling the relationship between completion rate and the 85 remaining features using linear regression (statsmodels. regression.linear_model). Results from the initial linear regression indicated the features explained a moderate amount of variation in completion rate (R^2_{adj} = 0.40, DF = 83, P <0.01). However, the model also triggered a multicollinearity warning, indicating possible unreliable parameter estimates and therefore precluding statistical inference based on this initial regression model. To reduce the influence of multicollinearity, I conducted feature selection using variance inflation factor (VIF) scores (statsmodels.stats.outliers_influence.variance_inflation_factor). Following general guidance regarding VIF scores, I removed continuous features with VIF scores > 5. This resulted in the removal of 34 features and left 51 features plus the target, such that the working DataFrame contained 19803 cases and 52 features including the target completion rate ('C150_L4').



Appendix Fig. 2. Autocorrelation of residuals plot (ACF) for the linear regression with a time lag of 40. The ACF plot shows small positive, irregular autocorrelation values between consecutive residuals. These results suggest the model meets the linear regression assumption of independence between residuals.

Regression Diagnostics

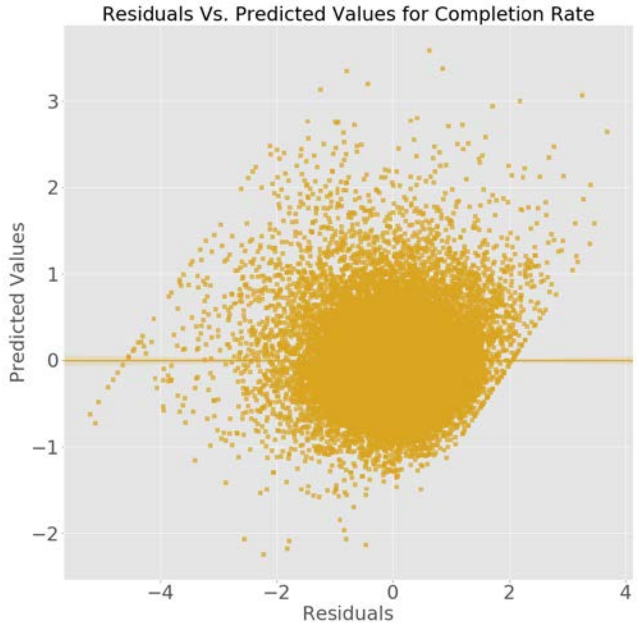
I used diagnostic tests to evaluate potential violation of the assumptions for linear regression. Violations of the linearity assumption lead to serious prediction errors, however, the linear pattern revealed by plotting observed and linearly predicted values implies meeting this assumption (Appendix Fig. 1).

Serially correlated errors mostly impact time-series data but indicate poor model specification and lead to systematic prediction errors. Results from a Durbin Watson test (DW = 1.85) and inspection of the autocorrelation function plot (Appendix Fig. 2) show scant evidence. For serial autocorrelation of residuals. Inconsistent variance (heteroskedasticity) of errors/residuals may result in confidence intervals that are too wide or too narrow or giving too much weight to a small subset of the data (namely the subset where the error variance was largest) when estimating coefficients. The Breusch Pagan test rejected the null hypothesis of homoscedasticity (LM = 1116.46, $P < 0.01$, F = 23.14, $P < 0.01$) and the double-outward box pattern created by plotting the residuals and predicted values (Appendix Fig. 3) supports the heteroskedasticity conclusion.

Finally, Violations of normality create problems for determining whether model coefficients are significantly different from zero and for calculating confidence intervals for forecasts. Results of tests evaluating this assumption were contradictory. The PDF of residuals appears Gaussian (Appendix Fig. 4) and the mean of residuals was near zero (0.005). However, the normal quantile plot suggests strong deviations from normality (Appendix Fig. 5), as did the normal test (scipy.stats.normaltest; statistic = 1776.89, P <0.01) and the Anderson-Darling test (statistic = 68.85, $P < 0.01$).

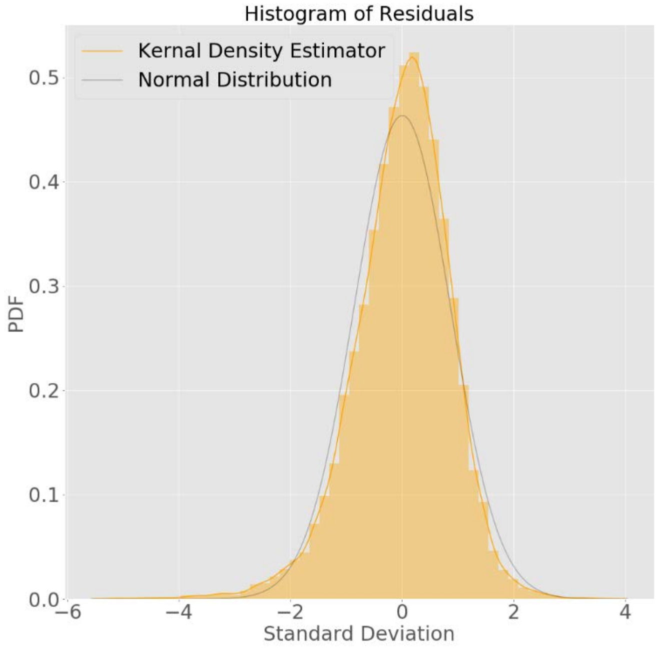
Linear Regression Model Diagnostics

Although the analyses provided useful insight, substantial room for improvement exists in statistical modeling and machine learning for explaining and predicting completion rate using the College Scorecard data. For example, the linear regression statistical model explained less than a third of the variation in completion rate. In addition, the machine learning models based on linear regression did not generalize well, based on cross validation for which the largest coefficient of determination for prediction was 0.20.

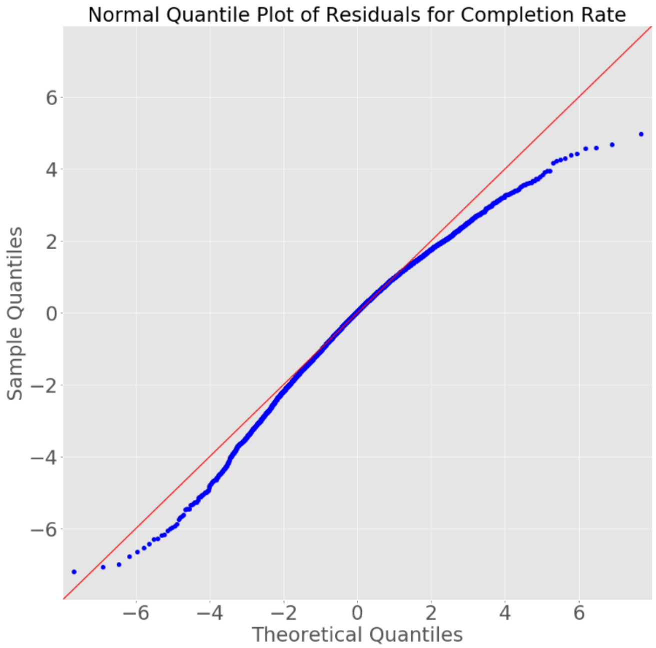


Appendix Fig. 3. Scatterplot comparing residuals and predicted values for the linear regression. The diamond-shape pattern displayed by the figure is also known as a "double outward box residuals distribution" which indicates heteroskedasticity and failure to meet the linear regression assumption of homoskedasticity.

Failure to adequately meet all the assumptions of linear regression, as illustrated by regression diagnostics helps explain the lackluster performance of the linear regression statistical model and machine learning regression algorithms. Regression diagnostics revealed heteroskedastic error variance and possible violations of the assumption of normally-distributed errors. Violations of these assumptions reduce accuracy of confidence intervals, create problems for determining statistical significance, and reduce a model's ability to extrapolate to unseen data.



Appendix Fig. 4. Histogram of residuals from the linear regression. Comparison the residuals with a super-imposed normal distribution (black line) indicates minor deviations from a Gaussian appearance.



Appendix Fig. 5. Normal quantile plot of residuals from linear regression compared to theoretical quantiles of a normal distribution. Deviation of the plotted residuals from the diagonal towards the extremes indicates the model fails to meet the linear regression assumption of normally-distributed residuals.

Appendix Table 1. Coefficients, descriptions and 95% confidence intervals generated from linear regression relating completion rate to the set of predictors.

Features	Coefficients	Description	95% CI Lower Bound	95% CI Upper Bound	P-value
PCIP01	0.162176	% degrees awarded in Agriculture, Agriculture Operations, And Related Sciences	0.14633	0.178021	< 0.01
PCIP46	0.125611	% degrees awarded in Construction Trades	0.111963	0.139259	< 0.01
PCIP48	0.099247	% degrees awarded in Precision Production	0.085584	0.112911	< 0.01
MAIN	0.096429	Main campus	0.076076	0.116781	< 0.01
INEXPFTE	0.092421	Instructional expenditures per full-time equivalent student	0.077569	0.107272	< 0.01
NUMBRANCH	0.084382	Number of branch campuses	0.064197	0.104568	< 0.01
VETERAN	0.067631	Share of veteran students	0.053726	0.081536	< 0.01
INC_PCT_M1	0.061921	Aided students with family incomes between $30,001-$48,000 in nominal dollars	0.042024	0.081818	< 0.01
PCIP09	0.060377	% degrees awarded in Communication, Journalism, And Related Programs	0.046933	0.073821	< 0.01
PCIP49	0.046801	% degrees awarded in Transportation And Materials Moving	0.034335	0.059267	< 0.01
PCIP10	0.045769	% degrees awarded in Communications Technologies/Technicians And Support Services	0.032929	0.058609	< 0.01
PCIP23	0.043922	% degrees awarded in English Language And Literature/Letters	0.030778	0.057066	< 0.01
PCIP30	0.043822	% degrees awarded in Multi/Interdisciplinary Studies	0.031268	0.056375	< 0.01
OPENADMP	0.041412	Non-open Admissions Policy	0.028121	0.054703	< 0.01
PCIP42	0.03868	% degrees awarded in Psychology	0.023521	0.053838	< 0.01
PCIP45	0.03534	% degrees awarded in Social Sciences	0.021869	0.048812	< 0.01
PCIP14	0.029081	% degrees awarded in Engineering	0.01634	0.041823	< 0.01

PCIP26	0.025455	% degrees awarded in Biological And Biomedical Sciences	0.012401	0.038508	< 0.01
PCIP27	0.024675	% degrees awarded in Mathematics And Statistics	0.010474	0.038875	< 0.01
PCIP43	0.023785	% degrees awarded in Homeland Security, Law Enforcement, Firefighting And Related Protective Services	0.011176	0.036394	< 0.01
MD_FAMINC	0.02369	Median family income in real 2015 dollars	0.003701	0.043679	0.02
UGDS_UNKN	0.021135	Total share of enrollment of undergraduate degree-seeking students whose race is unknown	0.008618	0.033652	< 0.01
PAR_ED_PCT_MS	0.019957	% students whose parents' highest educational level is middle school	0.004728	0.035187	0.01
PCIP38	0.008252	% degrees awarded in Philosophy And Religious Studies	-0.00345	0.019951	0.17
PCIP31	0.0068	% degrees awarded in Parks, Recreation, Leisure, And Fitness Studies	-0.00635	0.019952	0.31
PCIP25	0.004251	% degrees awarded in Library Science	-0.00935	0.017854	0.54
PCIP40	0.002403	% degrees awarded in Physical Sciences	-0.00949	0.014293	0.69
PCIP04	-0.00342	% degrees awarded in Architecture And Related Services	-0.01647	0.009627	0.61
PCIP12	-0.00615	% degrees awarded in Personal And Culinary Services	-0.01868	0.006373	0.34
PCIP39	-0.00638	% degrees awarded in Theology And Religious Vocations	-0.0182	0.005435	0.29
PCIP54	-0.00752	% degrees awarded in History	-0.02269	0.007662	0.33
AVGFACSAL	-0.00881	Average faculty salary	-0.02524	0.007619	0.29
PCIP29	-0.01558	% degrees awarded in Military Technologies And Applied Sciences	-0.02793	-0.00324	0.01
PCIP03	-0.01626	% degrees awarded in Natural Resources And Conservation	-0.03014	-0.00238	0.02

PCIP19	-0.02258	% degrees awarded in Family And Consumer Sciences/Human Sciences	-0.03553	-0.00962	< 0.01
PCIP11	-0.02282	% degrees awarded in Computer And Information Sciences And Support Services	-0.03556	-0.01009	< 0.01
PCIP16	-0.023	% degrees awarded in Foreign Languages, Literatures, And Linguistics	-0.03668	-0.00933	< 0.01
UGDS_NRA	-0.02492	Total share of enrollment of undergraduate degree-seeking students who are non-resident aliens	-0.03791	-0.01193	< 0.01
TUITFTE	-0.02811	Net tuition revenue per full-time equivalent student	-0.04572	-0.0105	< 0.01
PCIP41	-0.02823	% degrees awarded in Science Technologies/Technicians	-0.0407	-0.01576	< 0.01
PCIP05	-0.03291	% degrees awarded in Area, Ethnic, Cultural, Gender, And Group Studies	-0.04849	-0.01734	< 0.01
TUITIONFEE_OUT	-0.04158	Out-of-state tuition and fees	-0.05809	-0.02508	< 0.01
PCIP50	-0.04285	% degrees awarded in Visual And Performing Arts	-0.0555	-0.03021	< 0.01
PCIP44	-0.05103	% degrees awarded in Public Administration And Social Service Professions	-0.06414	-0.03792	< 0.01
PCIP22	-0.05456	% degrees awarded in Legal Professions And Studies	-0.06739	-0.04173	< 0.01
TUITIONFEE_IN	-0.07678	In-state tuition and fees	-0.09635	-0.05722	< 0.01
DEP_INC_PCT_M1	-0.08005	Dependent students with family incomes between $30,001-$48,000 in nominal dollars	-0.09764	-0.06245	< 0.01
PCIP13	-0.10377	% degrees awarded in Education	-0.11733	-0.0902	< 0.01
PAR_ED_PCT_HS	-0.11225	% students whose parents' highest educational level is high school	-0.12718	-0.09732	< 0.01
PREDDEG	-0.18529	Predominate undergraduate degree awarded is Associate's Degree	-0.19932	-0.17125	< 0.01
PPTUG_EF	-0.26662	Share of undergraduate, degree-/certificate-seeking students who are part-time	-0.28173	-0.25151	< 0.01