**Application of Machine Learning**

The initial data exploration has been useful to understand trends in the data and suggested what will be worth exploring in more detail in the regression models. The initial goal of the project as discussed in the Capstone Proposal was to create a linear regression model which will determine if the percentage of online students is significantly lower at HBCUs than at other institutions, after controlling for relevant factors. An additional goal was to explore whether, after controlling for relevant institutional characteristics, the addition of online courses leads to increased enrollments.

However, since the distribution of perc100online and perc\_some\_online are highly skewed, with most institutions having no students enrolled in fully online programs or taking online classes, one cannot use a linear regression model. One of the assumptions of linear regression is that the dependent variable is normally distributed.

Therefore, the initial goal the analysis was modified. A logistic regression model was used instead to identify the institutional characteristics that predict whether an institution offers online courses. Additionally, a logistic model was built to explain perc\_some\_online instead of perc100online. This is because the percentage of students in fully online programs is relatively low (7%) in the sample. This compares to 13% for the students who take some of their classes online. The variable perc\_some\_online was recoded as a categorical variable with two levels: 1 – offers some online classes; 2 no online classes. The logistic regression answers the following research question:

**RQ1. After controlling for size and other relevant institutional characteristics, are HBCUs significantly less likely than other institutions to offer online classes?**

In addition to the logistic model with perc\_some\_online as the dependent variable, a linear regression model was built to estimate whether there is a significant gain in total enrollments from offering online classes, after controlling for relevant institutional characteristics. This model answered the following research question:

**RQ2. After controlling for size and other relevant institutional characteristics, does offering online classes yield a significant increase in enrollments?**

For both models, the data set was split randomly into two data sets: a training data set consisting of 85% of the observations and a test data set with 15% of the observations. In terms machine learning methods, both models are considered to be supervised regression.

**Logistic Regression Model**

The results of the logistic model are presented below:

Call:

glm(formula = some\_online ~ CONTROL + MEDICAL + HOSPITAL + LANDGRNT +

UGOFFER + GROFFER + HBCU + EFDETOT + LOCALE + HLOFFER, family = binomial,

data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.0043 0.3496 0.4253 0.4843 1.1439

**Coefficients:**

**Estimate Std. Error z value Pr(>|z|)**

(Intercept) 0.6789021 1.8839132 0.360 0.71857

CONTROLpublic -0.2127811 0.3041801 -0.700 0.48422

MEDICALGrants medical degree 0.4737612 0.7893316 0.600 0.54837

HOSPITALNo Hospital 0.3798091 0.9030661 0.421 0.67406

LANDGRNTNot a Land Grant Institution 1.7945039 0.6476859 2.771 0.00559 \*\*

UGOFFERundergraduate degree 0.2907398 0.5293928 0.549 0.58287

GROFFERno graduate offering 0.1877876 0.4777077 0.393 0.69424

HBCUNot an HBCU -1.5881744 0.8048412 -1.973 0.04846 \*

EFDETOT 0.0003030 0.0001008 3.008 0.00263 \*\*

LOCALErural 0.1754421 0.3414386 0.514 0.60737

LOCALEsuburb -0.1669362 0.2757076 -0.605 0.54486

LOCALEtown -0.0986749 0.2680459 -0.368 0.71278

HLOFFER 0.0146464 0.1195157 0.123 0.90247

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 781.11 on 1247 degrees of freedom

Residual deviance: 757.66 on 1235 degrees of freedom

(75 observations deleted due to missingness)

AIC: 783.66

Number of Fisher Scoring iterations: 6

In order to interpret the results of the logistic regression, the coefficients were transformed. The new coefficients are presented in the following table:

**Estimate Std. Error z value Pr(>|z|)**

(Intercept) 0.6789020843 1.8839132294 0.3603680 0.718571937

CONTROLpublic -0.2127810965 0.3041800934 -0.6995234 0.484224988

MEDICALGrants medical degree 0.4737612144 0.7893316284 0.6002055 0.548369264

HOSPITALNo Hospital 0.3798090586 0.9030661387 0.4205772 0.674063821

LANDGRNTNot a Land Grant Institution 1.7945039311 0.6476859065 2.7706392 0.005594639

UGOFFERundergraduate degree 0.2907397868 0.5293927777 0.5491948 0.582871739

GROFFERno graduate offering 0.1877875621 0.4777076688 0.3931014 0.694244577

HBCUNot an HBCU -1.5881743771 0.8048411665 -1.9732768 0.048464038

EFDETOT 0.0003030417 0.0001007504 3.0078460 0.002631064

LOCALErural 0.1754420804 0.3414385641 0.5138321 0.607369412

LOCALEsuburb -0.1669362330 0.2757076417 -0.6054828 0.544858254

LOCALEtown -0.0986749333 0.2680458881 -0.3681270 0.712778526

HLOFFER 0.0146463907 0.1195156734 0.1225479 0.902465140

The results indicate that, after controlling for key institutional characteristics such as size, location, and highest degree offered, HBCUs are 1.6 times more likely than non HBCUs to have students enrolled in online classes. In addition, non land grant institutions are 1.8 time more likely than grant institutions to have students enrolled in online classes.

**Linear Regression Model**

A regression model was constructed to examine the impact of offering online classes on total enrollment.

Call:

lm(formula = EFDETOT ~ CONTROL + MEDICAL + HOSPITAL + LANDGRNT +

UGOFFER + GROFFER + HBCU + EFDETOT + LOCALE + HLOFFER + some\_online,

data = train)

Residuals:

Min 1Q Median 3Q Max

-2733.2 -762.6 -153.1 657.4 3401.3

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1732.507 585.115 -2.961 0.00313 \*\*

CONTROLpublic 1328.387 84.967 15.634 < 2e-16 \*\*\*

MEDICALGrants medical degree 5.438 229.513 0.024 0.98110

HOSPITALNo Hospital 250.980 281.540 0.891 0.37286

LANDGRNTNot a Land Grant Institution 390.944 259.801 1.505 0.13263

UGOFFERundergraduate degree 1184.962 182.583 6.490 1.24e-10 \*\*\*

GROFFERno graduate offering 222.670 147.730 1.507 0.13199

HBCUNot an HBCU 260.899 157.283 1.659 0.09741 .

LOCALErural -591.555 98.955 -5.978 2.95e-09 \*\*\*

LOCALEsuburb 6.131 84.885 0.072 0.94244

LOCALEtown -407.384 80.917 -5.035 5.50e-07 \*\*\*

HLOFFER 208.008 35.823 5.807 8.10e-09 \*\*\*

some\_onlinesome online enrollment 321.967 102.695 3.135 0.00176 \*\*

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1055 on 1235 degrees of freedom

(75 observations deleted due to missingness)

Multiple R-squared: 0.2068, Adjusted R-squared: 0.1991

F-statistic: 26.83 on 12 and 1235 DF, p-value: < 2.2e-16

The model explains 20% of the variance in total enrollment. The results suggest that, on average, after controlling for relevant institutional characteristics, offering online classes results in 322 additional students. Offering undergraduate programs adds on average 1185 students. Public institutions have on average 1328 more students than private institutions. In addition, compared to institutions located in large urban areas, institutions located in towns enroll on average 407 fewer students. Institutions located in rural areas have 592 fewer students on average than urban institutions. Lastly, the higher the degree offered by the institution, the higher the total enrollment.