Predicting Earnings after Graduation

Jenna Kutz, Percy Musiiwa, Matt Jackson

March 12, 2024

Problem Statement

Education is the cornerstone of a successful career. In a fast-moving society, the vast majority of potential employers seek to hire individuals with postsecondary degrees and certificates.

According to the Association of Public & Land-Grant Universities,

> *"The evidence that a college degree significantly improves one's employment prospects and earnings potential is overwhelming… On an annual basis, median earnings for bachelor's degree holders are $36,000 or 84 percent higher than those whose highest degree is a high school diploma."*

As a cohort that is soon to complete a six-month certificate, we (at least intuitively) understand the value of education. But how can a potential student be more deliberate if the goal is to maximize income over the course of a career?

Source: aplu.org

Project Overview

Goals:

- Provide insights on post-graduate earnings from the data.gov college scorecard
- Create a multiple linear regression model that will provide insight into the universities and features that correlate to higher earnings
- Utilize tools from previous modules to complete the analysis:
  - PySpark / SQL
  - Data cleansing, standardization tools (Standard Scaler)
  - Linear regression libraries (sklearn)
  - General data analysis tools (Python, Pandas, etc.)

Data Exploration and Familiarization

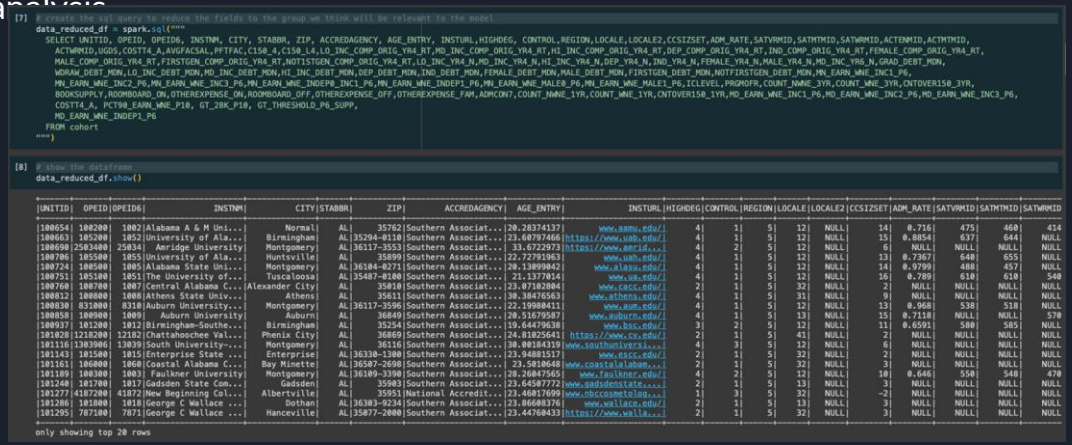
- Dataset contained ~6500 rows and ~2900 columns
- A data dictionary was included (very helpful!)
- Data categorized into broad categories:
  - Academics
  - Admissions
  - Aid
  - Completion
  - Cost
  - Earnings
  - Repayment
  - School
  - Student

For our analysis, we tried to build as much familiarity with the cost, earnings, school and student domains.

Data Prep and Ingestion

The data.gov College Scorecard was broken into multiple files, each ~250MB to prep the data for ingestion we took these steps:

- Created a zip file (~24MB) of the most recent cohort that can be loaded to our GitHub repo
- Aligned on the use of Google Colab as our notebook environment
- Aligned on using Spark to ingest and perform the initial rendering of the data
- Collectively selected ~80 fields that would serve as our starting point for feature analysis



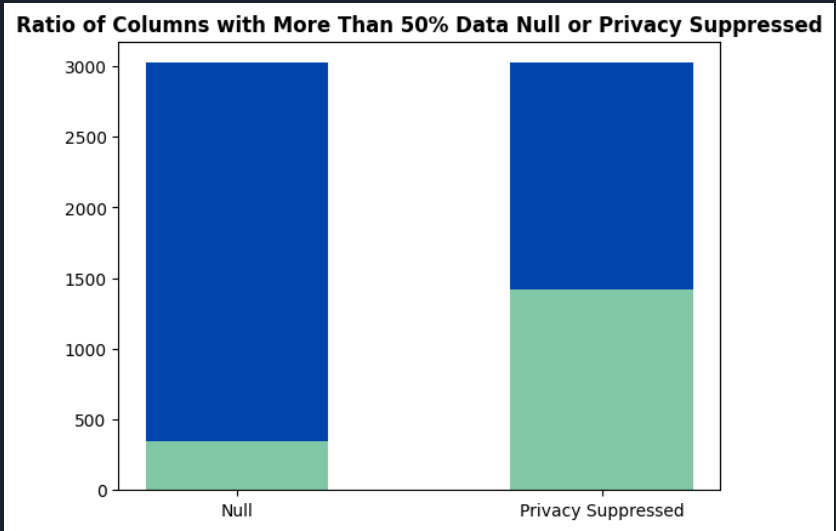
```
[7] # create the sql query to reduce the fields to the group we think will be relevant to the model
data_reduced_df = spark.sql("""
  SELECT UNITID, OPEID, OPEID6, INSTNM, CITY, STABBR, ZIP, ACCREDAGENCY, AGE_ENTRY, INSTURL,HIGHDEG, CONTROL,REGION,LOCALE,LOCALE2,CCSIZSET,ADM_RATE,SATVRMID,SATMTMID,SATWRMID,ACTENMID,ACTMTMID,
    ACTWRMID,UGDS,COSTT4_A,AVGFACSAL,PFTFAC,C150_4,C150_L4,LO_INC_COMP_ORIG_YR4_RT,MD_INC_COMP_ORIG_YR4_RT,HI_INC_COMP_ORIG_YR4_RT,DEP_COMP_ORIG_YR4_RT,IND_COMP_ORIG_YR4_RT,FEMALE_COMP_ORIG_YR4_RT,
    MALE_COMP_ORIG_YR4_RT,FIRSTGEN_COMP_ORIG_YR4_RT,NOT1STGEN_COMP_ORIG_YR4_RT,LO_INC_YR4_N,MD_INC_YR4_N,HI_INC_YR4_N,DEP_YR4_N,IND_YR4_N,FEMALE_YR4_N,MALE_YR4_N,MD_INC_YRS_N,GRAD_DEBT_MDN,
    WDRAW_DEBT_MDN,LO_INC_DEBT_MDN,MD_INC_DEBT_MDN,HI_INC_DEBT_MDN,DEP_DEBT_MDN,IND_DEBT_MDN,FEMALE_DEBT_MDN,MALE_DEBT_MDN,FIRSTGEN_DEBT_MDN,NOTFIRSTGEN_DEBT_MDN,MN_EARN_WNE_INC1_P6,
    MN_EARN_WNE_INC2_P6,MN_EARN_WNE_INC3_P6,MN_EARN_WNE_INDEP0_INC1_P6,MN_EARN_WNE_INDEP1_P6,MN_EARN_WNE_MALE0_P6,MN_EARN_WNE_MALE1_P6,ICLEVEL,PRGMOFR,COUNT_NWNE_3YR,COUNT_WNE_3YR,CNTOVER150_3YR,
    BOOKSUPPLY,ROOMBOARD_ON,OTHEREXPENSE_ON,ROOMBOARD_OFF,OTHEREXPENSE_OFF,OTHEREXPENSE_FAM,ADMCONT7,COUNT_NWNE_1YR,COUNT_WNE_1YR,CNTOVER150_1YR,MD_EARN_WNE_INC1_P6,MD_EARN_WNE_INC2_P6,MD_EARN_WNE_INC3_P6,
    COSTT4_A, PCT90_EARN_WNE_P10, GT_28K_P10, GT_THRESHOLD_P6_SUPP,
    MD_EARN_WNE_INDEP1_P6
  FROM cohort
  """)

[8] # show the dataframe
data_reduced_df.show()
```

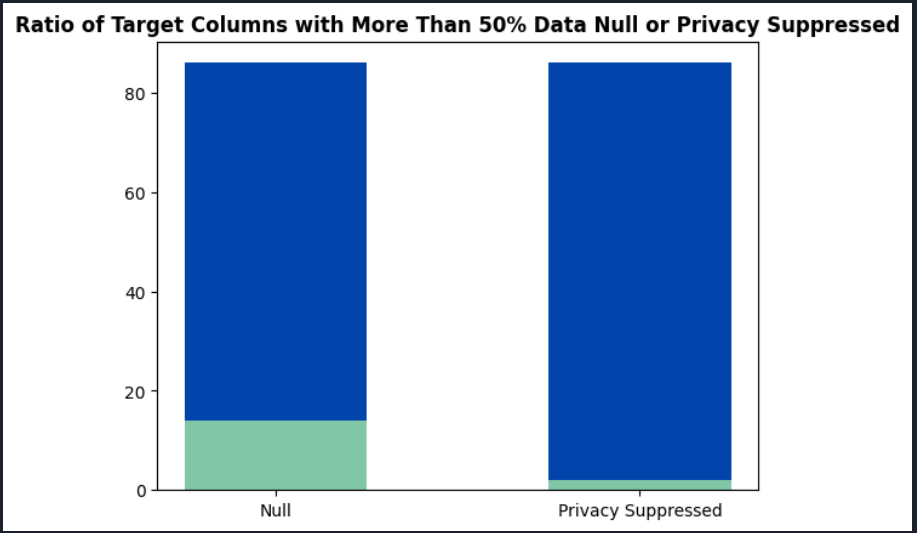
UNITID	OPEID	OPEID6	INSTNM	CITY	STABBR	ZIP	ACCREDAGENCY	AGE_ENTRY	INSTURL	HIGHDEG	CONTROL	REGION	LOCALE	LOCALE2	CCSIZSET	ADM_RATE	SATVRMID	SATMTMID	SATWRMID
100654	100200	1002	Alabama A & M Uni...	Normal	AL	35762	Southern Associat...	20.28374137	www.aamu.edu/	4	1	5	12	NULL	14	0.716	475	460	414
100663	105200	1052	University of Ala...	Birmingham	AL	35294-0110	Southern Associat...	23.60797466	https://www.uab.edu/	4	1	5	12	NULL	15	0.8054	637	644	NULL
100690	2503400	25034	Amridge University	Montgomery	AL	36117-3553	Southern Associat...	33.6722973	https://www.amrid...	4	2	5	12	NULL	6	NULL	NULL	NULL	NULL
100706	105500	1055	University of Ala...	Huntsville	AL	35899	Southern Associat...	22.72791963	www.uah.edu/	4	1	5	12	NULL	13	0.7367	640	655	NULL
100724	100500	1005	Alabama State Uni...	Montgomery	AL	36104-0271	Southern Associat...	20.13099042	www.alasu.edu/	4	1	5	12	NULL	14	0.9799	488	457	NULL
100751	105100	1051	The University of...	Tuscaloosa	AL	35487-0100	Southern Associat...	21.1377014	www.ua.edu/	4	1	5	12	NULL	16	0.789	630	610	540
100760	100700	1007	Central Alabama C...	Alexander City	AL	35010	Southern Associat...	23.07102004	www.cacc.edu/	2	1	5	32	NULL	2	NULL	NULL	NULL	NULL
100812	100800	1008	Athens State Univ...	Athens	AL	35611	Southern Associat...	30.38476563	www.athens.edu/	4	1	5	31	NULL	9	NULL	NULL	NULL	NULL
100830	831800	8318	Auburn University...	Montgomery	AL	36117-3596	Southern Associat...	22.35980411	www.aum.edu/	4	1	5	13	NULL	13	0.968	530	510	NULL
100858	100900	1009	Auburn University	Auburn	AL	36849	Southern Associat...	20.51679587	www.auburn.edu/	4	1	5	13	NULL	15	0.7118	NULL	NULL	570
100937	101200	1012	Birmingham-Southe...	Birmingham	AL	35254	Southern Associat...	19.64479638	www.bsc.edu/	3	2	5	12	NULL	11	0.6591	580	585	NULL
101028	1218200	12182	Chattahoochee Val...	Phenix City	AL	36869	Southern Associat...	24.81025641	https://www.cv.edu/	2	1	5	41	NULL	2	NULL	NULL	NULL	NULL
101116	1303900	13039	South University-...	Montgomery	AL	36116	Southern Associat...	30.00184319	www.southuniversi...	4	3	5	12	NULL	6	NULL	NULL	NULL	NULL
101143	101500	1015	Enterprise State ...	Enterprise	AL	36330-1300	Southern Associat...	23.94881517	www.escc.edu/	2	1	5	32	NULL	2	NULL	NULL	NULL	NULL
101161	106000	1060	Coastal Alabama C...	Bay Minette	AL	36507-2698	Southern Associat...	23.5010648	www.coastalalabam...	2	1	5	32	NULL	3	NULL	NULL	NULL	NULL
101189	100300	1003	Faulkner University	Montgomery	AL	36109-3390	Southern Associat...	20.26047565	www.faulkner.edu/	4	2	5	12	NULL	10	0.646	550	540	470
101240	101700	1017	Gadsden State Com...	Gadsden	AL	35903	Southern Associat...	23.64507772	www.gadsdenstate....	2	1	5	13	NULL	3	NULL	NULL	NULL	NULL
101277	4187200	41872	New Beginning Col...	Albertville	AL	35951	National Accredit...	23.46017699	www.nbccosmetolog...	1	3	5	32	NULL	-2	NULL	NULL	NULL	NULL
101286	101800	1018	George C Wallace ...	Dothan	AL	36303-9234	Southern Associat...	23.86608376	www.wallace.edu/	2	1	5	13	NULL	3	NULL	NULL	NULL	NULL
101295	787100	7871	George C Wallace ...	Hanceville	AL	35077-2000	Southern Associat...	23.44760433	https://www.walla...	2	1	5	32	NULL	3	NULL	NULL	NULL	NULL

only showing top 20 rows

Data Cleanup & Feature Selection

At this point, we decided that we would focus on creating a multiple linear regression model. To prep the data for this type of analysis, we'd need to remove string and null values from the





Target variable:

Median earnings of independent students working and not enrolled 6 years after entry

Model Training and Optimization

- After converting the Spark dataframe to a Pandas dataframe, we ran the .corr() method on the remaining features to gather a baseline for relatedness to our target variable.
- After importing the sklearn modules, we ran the initial data.
- At first, our model was only performing in the 55-65% range. We had to find more feature variables that were better correlated to the target.
- Optimizing the model took approximately 10-15 attempts but we were able to get it above 80%

Correlation to Target

```
model_fields.corr()['MD_EARN_WNE_INDEP1_P6']
```

```
HIGHDEG                          0.524405
AGE_ENTRY                       -0.231888
NOTFIRSTGEN_DEBT_MDN             0.559862
FIRSTGEN_DEBT_MDN                0.571523
MALE_DEBT_MDN                    0.601051
FEMALE_DEBT_MDN                  0.549165
IND_DEBT_MDN                     0.500618
DEP_DEBT_MDN                     0.592892
HI_INC_DEBT_MDN                  0.557926
MD_INC_DEBT_MDN                  0.552280
LO_INC_DEBT_MDN                  0.544745
WDRAW_DEBT_MDN                   0.421478
GRAD_DEBT_MDN                    0.376582
MD_INC_YR6_N                     0.002820
MALE_YR4_N                       0.007979
FEMALE_YR4_N                    -0.010155
IND_YR4_N                       -0.071919
DEP_YR4_N                        0.094483
HI_INC_YR4_N                     0.301812
MD_INC_YR4_N                     0.057178
LO_INC_YR4_N                    -0.150689
NOT1STGEN_COMP_ORIG_YR4_RT       0.595411
FIRSTGEN_COMP_ORIG_YR4_RT        0.624209
MALE_COMP_ORIG_YR4_RT            0.621164
FEMALE_COMP_ORIG_YR4_RT          0.620534
IND_COMP_ORIG_YR4_RT             0.622742
DEP_COMP_ORIG_YR4_RT             0.589223
HI_INC_COMP_ORIG_YR4_RT          0.503009
MD_INC_COMP_ORIG_YR4_RT          0.544615
LO_INC_COMP_ORIG_YR4_RT          0.640177
COUNT_NWNE_1YR                   0.036239
COUNT_NWNE_3YR                  -0.058764
CCSIZSET                         0.512436
UGDS                             0.152350
COSTT4_A                         0.549476
PCT90_EARN_WNE_P10               0.733801
GT_28K_P10                       0.803089
GT_THRESHOLD_P6_SUPP             0.874590
MD_EARN_WNE_INDEP1_P6            1.000000
Name: MD_EARN_WNE_INDEP1_P6, dtype: float64
```

Output and Conclusions

```
# Compute metrics for the linear regression model: score, r2, mse, rmse, std
score = model.score(X, y, sample_weight=None)
mse = mean_squared_error(y_test, predictions)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, predictions)
# std = np.std(y)

# Print relevant metrics.
print(f"The score is {score}.")
print(f"The mean squared error is {mse}.")
print(f"The root mean squared error is {rmse}.")
print(f"The r2 score is {r2}.")
# print(f"The standard deviation is {std}.")
```

```
The score is -719897458.3919673.
The mean squared error is 24271349.987317916.
The root mean squared error is 4926.5961867518545.
The r2 score is 0.8330727452894079.
```