PREDICTING COLLEGE GRADUATION BASED ON HIGH SCHOOL MATH

BY,

KAS EVSEEV

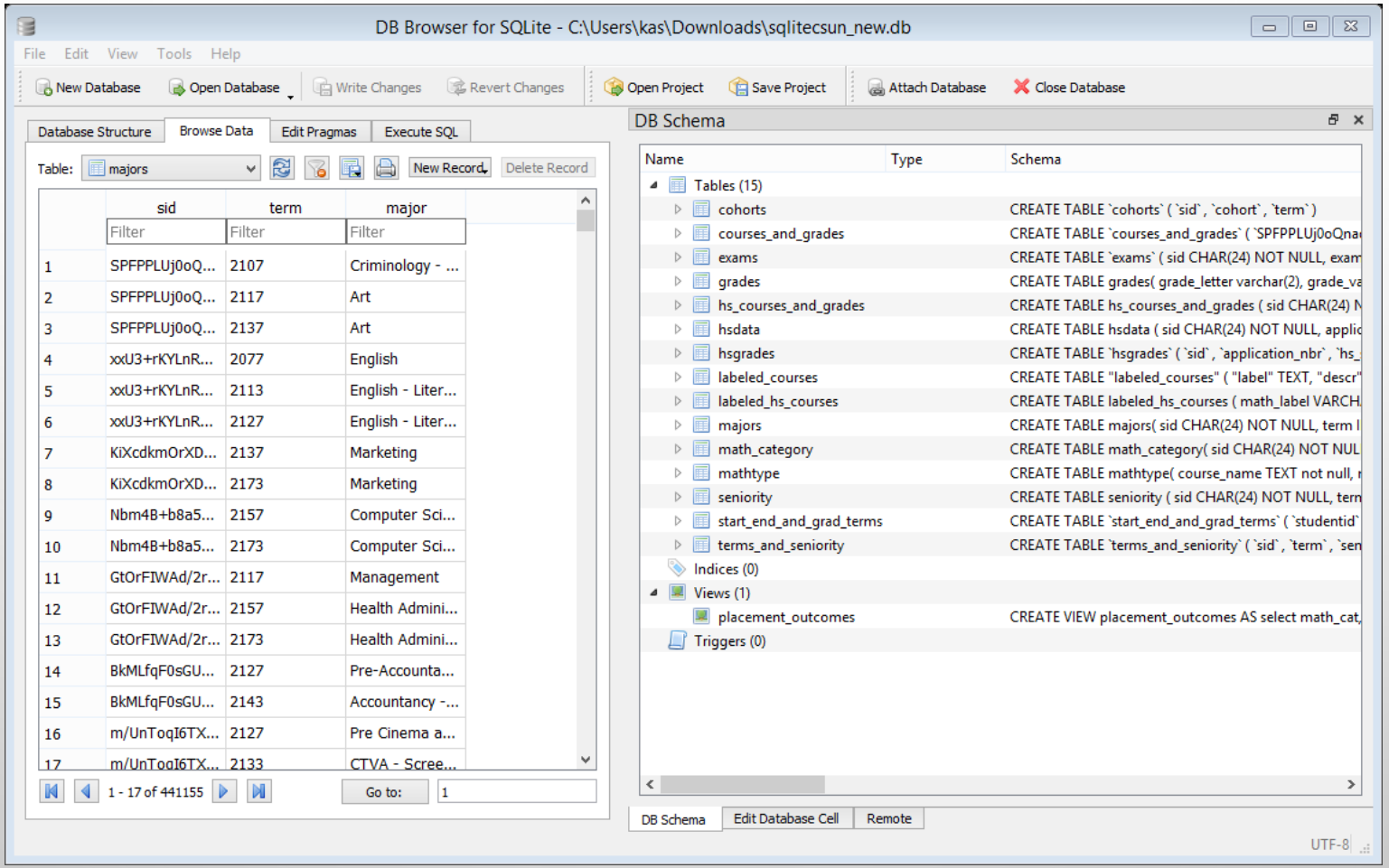
CONTENTS

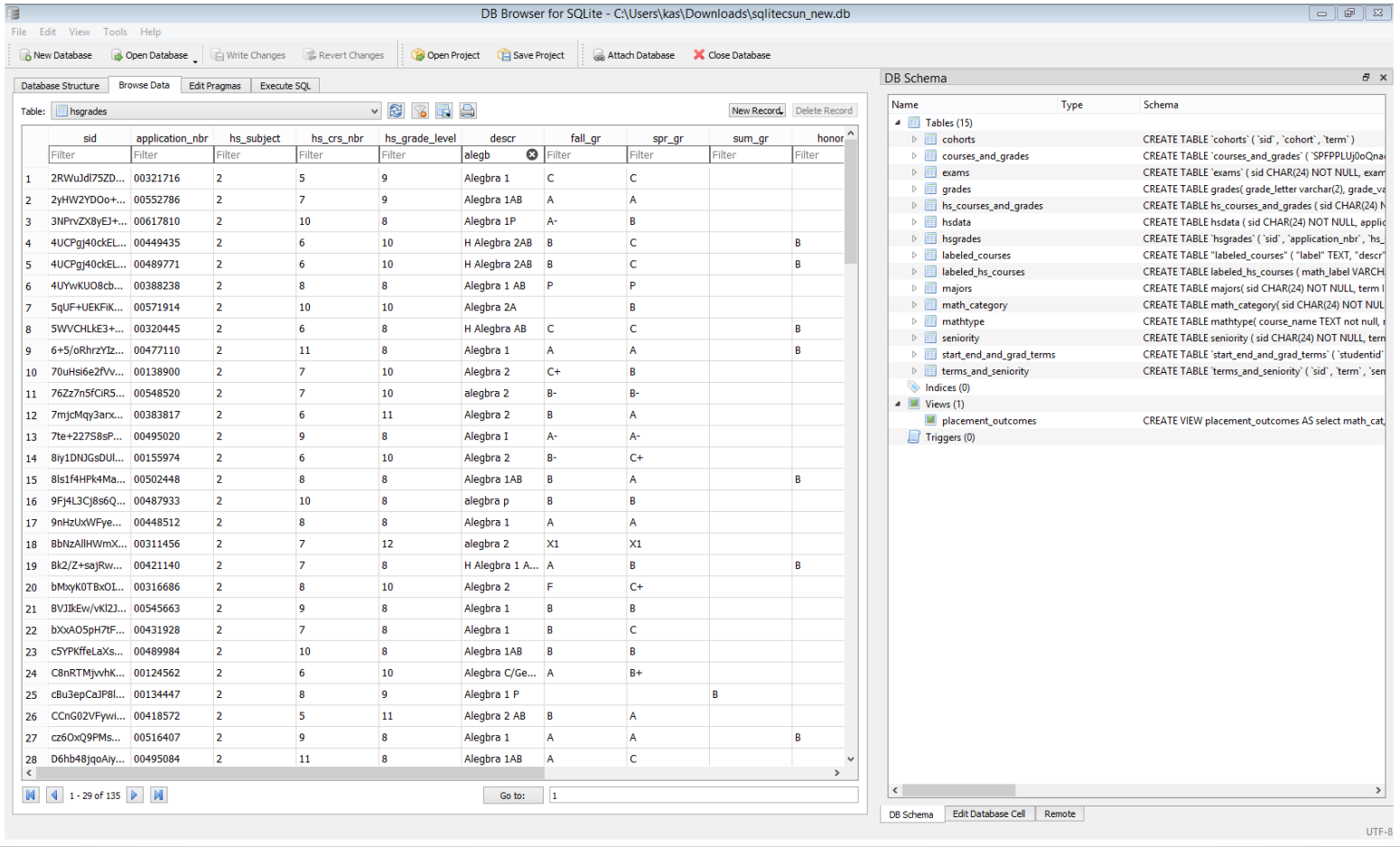
PROJECT DESCRIPTION

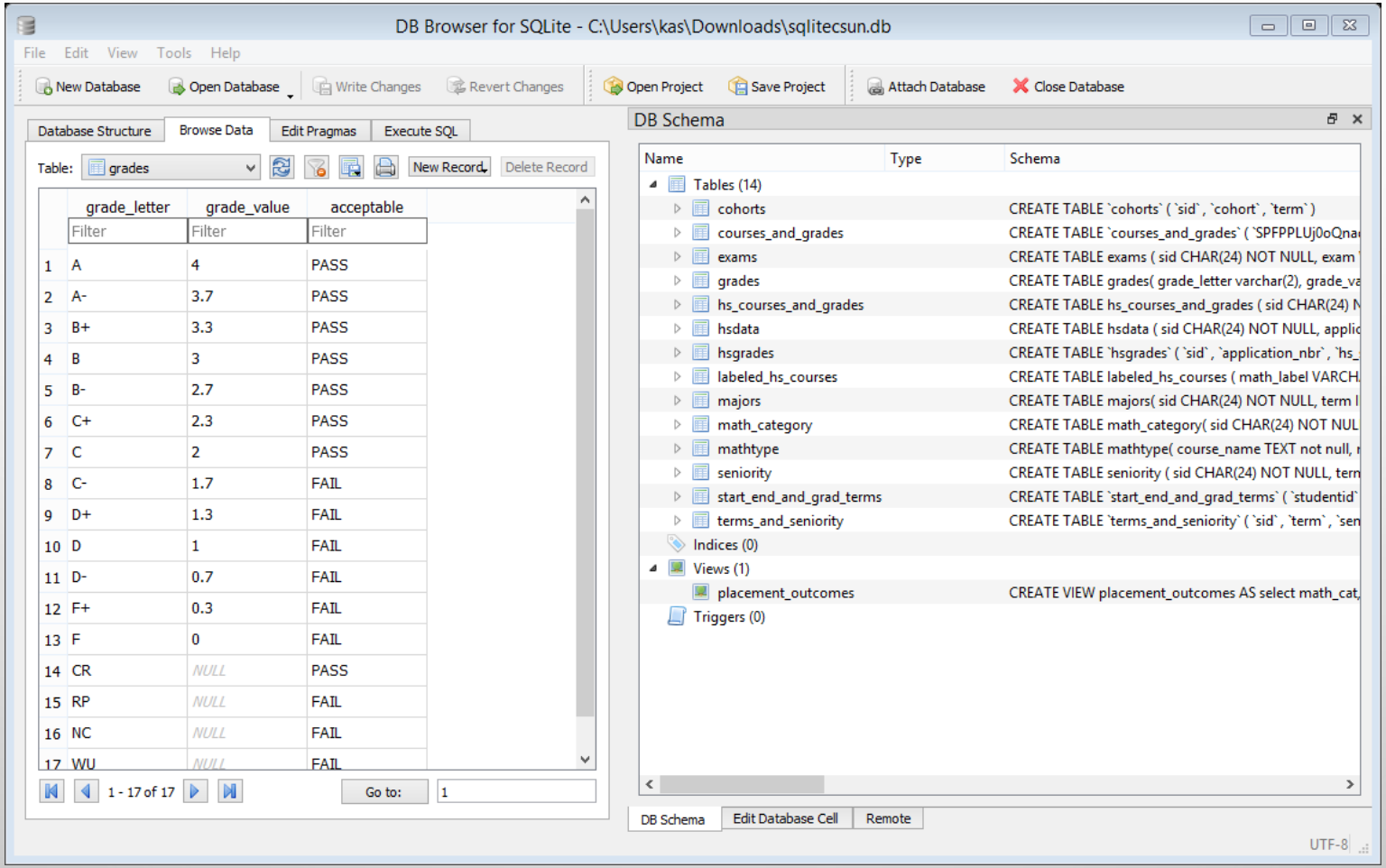
- WOULD GRADUATION RATES IMPROVE BY REQUIRING APPLICANTS TO TAKE 12TH GRADE MATH?
- ASSUME THEY WOULD NOT
- WANT TO FIND A WAY TO PREDICT WHO WILL GRADUATE
- IS HIGH SCHOOL MATH A GOOD PREDICTOR? BETTER THAN SAT SCORE?
- DATA ACQUIRED DATA ABOUT CSUN STUDENTS FROM CSU MAIN OFFICE
  - UNIQUE SOCIO-ETHNIC MAKEUP AT CSUN COMPARED TO ALL CSU CAMPUSES
- COMPARE ACCURACIES OF PREDICTIONS BASED ON HS MATH VS ON SAT SCORES

OVERVIEW OF DATA

- DATABASE COMES FROM CSU MAIN OFFICE
- FROM 2004 TO 2017
- ANONYMIZED
- MULTIPLE TABLES
- HIGH SCHOOL AND COLLEGE INFORMATION


DB Browser for SQLite - C:\Users\kas\Downloads\sqlitecsun_new.db
File Edit View Tools Help
New Database
Open Database
Write Changes
Revert Changes
Open Project
Save Project
Attach Database
Close Database
Database Structure
Browse Data
Edit Pragmas
Execute SQL
Table: majors
New Record
Delete Record
sid
term
major
Filter
1 SPFPPLUj0oQ... 2107 Criminology - ...
2 SPFPPLUj0oQ... 2117 Art
3 SPFPPLUj0oQ... 2137 Art
4 xxU3+rKYLnR... 2077 English
5 xxU3+rKYLnR... 2113 English - Liter...
6 xxU3+rKYLnR... 2127 English - Liter...
7 KiXcdkmOrXD... 2137 Marketing
8 KiXcdkmOrXD... 2173 Marketing
9 Nbm4B+b8a5... 2157 Computer Sci...
10 Nbm4B+b8a5... 2173 Computer Sci...
11 GtOrFIWAd/2r... 2117 Management
12 GtOrFIWAd/2r... 2157 Health Admini...
13 GtOrFIWAd/2r... 2173 Health Admini...
14 BkMLfqF0sGU... 2127 Pre-Accounta...
15 BkMLfqF0sGU... 2143 Accountancy -...
16 m/UnToqI6TX... 2127 Pre Cinema a...
17 m/UnToqI6TX... 2133 CTVA - Scree...
1 - 17 of 441155
Go to: 1
DB Schema
Name
Type
Schema
Tables (15)
cohorts CREATE TABLE `cohorts` (`sid`, `cohort`, `term`)
courses_and_grades CREATE TABLE `courses_and_grades` (`SPFPPLUj0oQna
exams CREATE TABLE `exams` (sid CHAR(24) NOT NULL, exam
grades CREATE TABLE grades(grade_letter varchar(2), grade_va
hs_courses_and_grades CREATE TABLE hs_courses_and_grades (sid CHAR(24) N
hsdata CREATE TABLE hsdata (sid CHAR(24) NOT NULL, applic
hsgrades CREATE TABLE `hsgrades` (`sid`, `application_nbr`, `hs_
labeled_courses CREATE TABLE "labeled_courses" ("label" TEXT, "descr"
labeled_hs_courses CREATE TABLE labeled_hs_courses (math_label VARCH
majors CREATE TABLE majors(sid CHAR(24) NOT NULL, term I
math_category CREATE TABLE math_category(sid CHAR(24) NOT NUL
mathtype CREATE TABLE mathtype(course_name TEXT not null,
seniority CREATE TABLE seniority (sid CHAR(24) NOT NULL, tern
start_end_and_grad_terms CREATE TABLE `start_end_and_grad_terms` (`studentid`
terms_and_seniority CREATE TABLE `terms_and_seniority` (`sid`, `term`, `sen
Indices (0)
Views (1)
placement_outcomes CREATE VIEW placement_outcomes AS select math_cat,
Triggers (0)
DB Schema
Edit Database Cell
Remote
UTF-8


DB Browser for SQLite - C:\Users\kas\Downloads\sqlitecsun_new.db
File Edit View Tools Help
New Database
Open Database
Write Changes
Revert Changes
Open Project
Save Project
Attach Database
Close Database
Database Structure
Browse Data
Edit Pragmas
Execute SQL
Table: hsgrades
New Record
Delete Record
sid application_nbr hs_subject hs_crs_nbr hs_grade_level descr fall_gr spr_gr sum_gr honor
Filter
alegb
1 2RWuJdl75ZD... 00321716 2 5 9 Alegbra 1 C C
2 2yHW2YDOo+... 00552786 2 7 9 Alegbra 1AB A A
3 3NPrvZX8yEJ+... 00617810 2 10 8 Alegbra 1P A- B
4 4UCPgJ40ckEL... 00449435 2 6 10 H Alegbra 2AB B C B
5 4UCPgJ40ckEL... 00489771 2 6 10 H Alegbra 2AB B C B
6 4UYwKUO8cb... 00388238 2 8 8 Alegbra 1 AB P P
7 5qUF+UEKFiK... 00571914 2 10 10 Alegbra 2A B
8 5WVCHLkE3+... 00320445 2 6 8 H Alegbra AB C C B
9 6+5/oRhrzY1z... 00477110 2 11 8 Alegbra 1 A A B
10 70uHsi6e2fVv... 00138900 2 7 10 Alegbra 2 C+ B
11 76Zz7n5fCiR5... 00548520 2 7 10 alegbra 2 B- B-
12 7mjcMqy3arx... 00383817 2 6 11 Alegbra 2 B A
13 7te+227S8sP... 00495020 2 9 8 Alegbra I A- A-
14 8iy1DNJGsDUl... 00155974 2 6 10 Alegbra 2 B- C+
15 8ls1f4HPk4Ma... 00502448 2 8 8 Alegbra 1AB B A B
16 9Fj4L3Cj8s6Q... 00487933 2 10 8 alegbra p B B
17 9nHzUxWFye... 00448512 2 8 8 Alegbra 1 A A
18 BbNzAllHWmX... 00311456 2 7 12 alegbra 2 X1 X1
19 Bk2/Z+sajRw... 00421140 2 7 8 H Alegbra 1 A... A B B
20 bMxyK0TBxOI... 00316686 2 8 10 Alegbra 2 F C+
21 BVJIkEw/vKl2J... 00545663 2 9 8 Alegbra 1 B B
22 bXxAO5pH7tF... 00431928 2 7 8 Alegbra 1 B C
23 c5YPKffeLaXs... 00489984 2 10 8 Alegbra 1AB B B
24 C8nRTMjvvhK... 00124562 2 6 10 Alegbra C/Ge... A B+
25 cBu3epCaJP8l... 00134447 2 8 9 Alegbra 1 P B
26 CCnG02VFywi... 00418572 2 5 11 Alegbra 2 AB B A
27 cz6OxQ9PMs... 00516407 2 9 8 Alegbra 1 A A B
28 D6hb48jqoAiy... 00495084 2 11 8 Alegbra 1AB A C
1 - 29 of 135
Go to: 1
DB Schema
Name Type Schema
Tables (15)
cohorts CREATE TABLE `cohorts` (`sid` , `cohort` , `term`)
courses_and_grades CREATE TABLE `courses_and_grades` (`SPFPPLUj0oQna
exams CREATE TABLE `exams` (sid CHAR(24) NOT NULL, exam
grades CREATE TABLE grades(grade_letter varchar(2), grade_va
hs_courses_and_grades CREATE TABLE hs_courses_and_grades (sid CHAR(24) N
hsdata CREATE TABLE hsdata (sid CHAR(24) NOT NULL, applic
hsgrades CREATE TABLE `hsgrades` (`sid` , `application_nbr` , `hs_
labeled_courses CREATE TABLE "labeled_courses" ("label" TEXT, "descr"
labeled_hs_courses CREATE TABLE labeled_hs_courses (math_label VARCH
majors CREATE TABLE majors(sid CHAR(24) NOT NULL, term I
math_category CREATE TABLE math_category(sid CHAR(24) NOT NUL
mathtype CREATE TABLE mathtype(course_name TEXT not null,
seniority CREATE TABLE seniority (sid CHAR(24) NOT NULL, tern
start_end_and_grad_terms CREATE TABLE `start_end_and_grad_terms` (`studentid`
terms_and_seniority CREATE TABLE `terms_and_seniority` (`sid` , `term` , `sen
Indices (0)
Views (1)
placement_outcomes CREATE VIEW placement_outcomes AS select math_cat
Triggers (0)
DB Schema
Edit Database Cell
Remote
UTF-8


DB Browser for SQLite - C:\Users\kas\Downloads\sqlitecsun.db
File Edit View Tools Help
New Database
Open Database
Write Changes
Revert Changes
Open Project
Save Project
Attach Database
Close Database
Database Structure
Browse Data
Edit Pragmas
Execute SQL
Table: grades
New Record
Delete Record
grade_letter grade_value acceptable
Filter Filter Filter
1 A 4 PASS
2 A- 3.7 PASS
3 B+ 3.3 PASS
4 B 3 PASS
5 B- 2.7 PASS
6 C+ 2.3 PASS
7 C 2 PASS
8 C- 1.7 FAIL
9 D+ 1.3 FAIL
10 D 1 FAIL
11 D- 0.7 FAIL
12 F+ 0.3 FAIL
13 F 0 FAIL
14 CR NULL PASS
15 RP NULL FAIL
16 NC NULL FAIL
17 WU NULL FAIL
1 - 17 of 17
Go to: 1
DB Schema
Name Type Schema
Tables (14)
cohorts CREATE TABLE `cohorts` (`sid` , `cohort` , `term`)
courses_and_grades CREATE TABLE `courses_and_grades` (`SPFPPLUj0oQna
exams CREATE TABLE exams (sid CHAR(24) NOT NULL, exam
grades CREATE TABLE grades(grade_letter varchar(2), grade_va
hs_courses_and_grades CREATE TABLE hs_courses_and_grades (sid CHAR(24) N
hsdata CREATE TABLE hsdata (sid CHAR(24) NOT NULL, applic
hsgrades CREATE TABLE `hsgrades` (`sid` , `application_nbr` , `hs_
labeled_hs_courses CREATE TABLE labeled_hs_courses (math_label VARCH
majors CREATE TABLE majors(sid CHAR(24) NOT NULL, term I
math_category CREATE TABLE math_category(sid CHAR(24) NOT NUL
mathtype CREATE TABLE mathtype(course_name TEXT not null, r
seniority CREATE TABLE seniority (sid CHAR(24) NOT NULL, tern
start_end_and_grad_terms CREATE TABLE `start_end_and_grad_terms` (`studentid`
terms_and_seniority CREATE TABLE `terms_and_seniority` (`sid` , `term` , `sen
Indices (0)
Views (1)
placement_outcomes CREATE VIEW placement_outcomes AS select math_cat,
Triggers (0)
DB Schema
Edit Database Cell
Remote
UTF-8

RESULTS

- LOGISTIC REGRESSION: DEFINITION

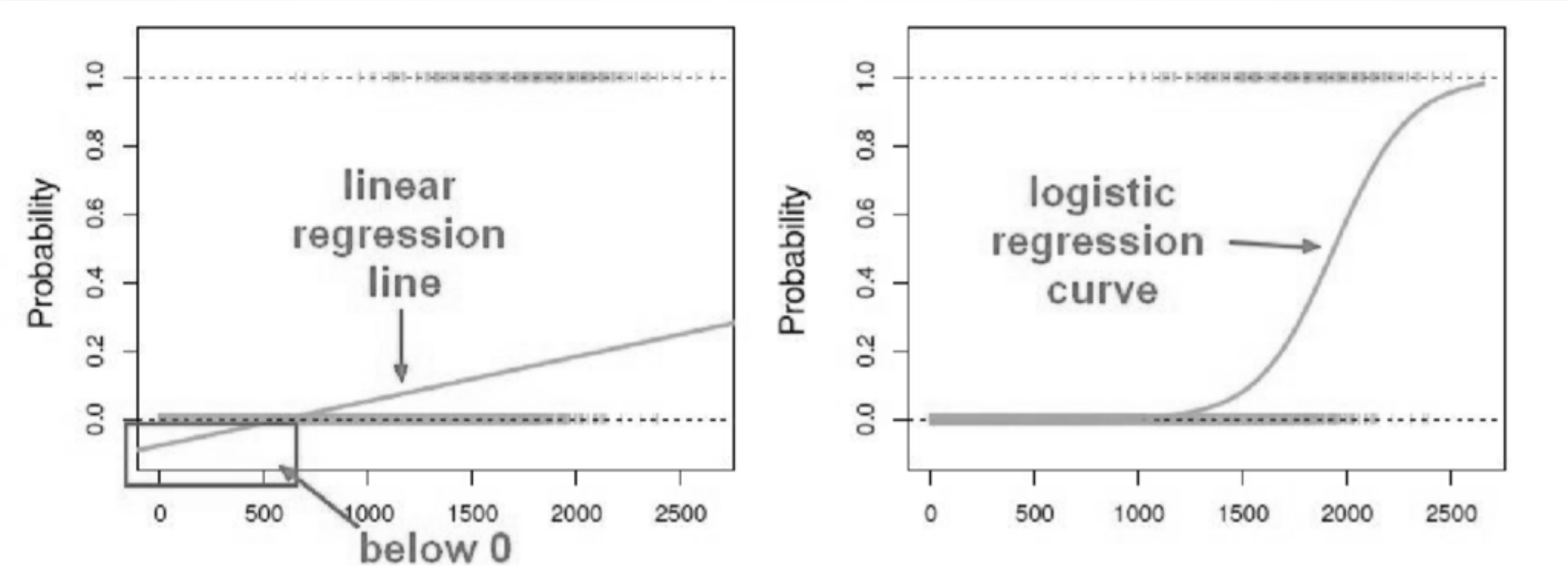
$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

We can recover the odds by exponentiating the log-odds:

$$\frac{p}{1-p} = b^{\beta_0+\beta_1 x_1+\beta_2 x_2}.$$

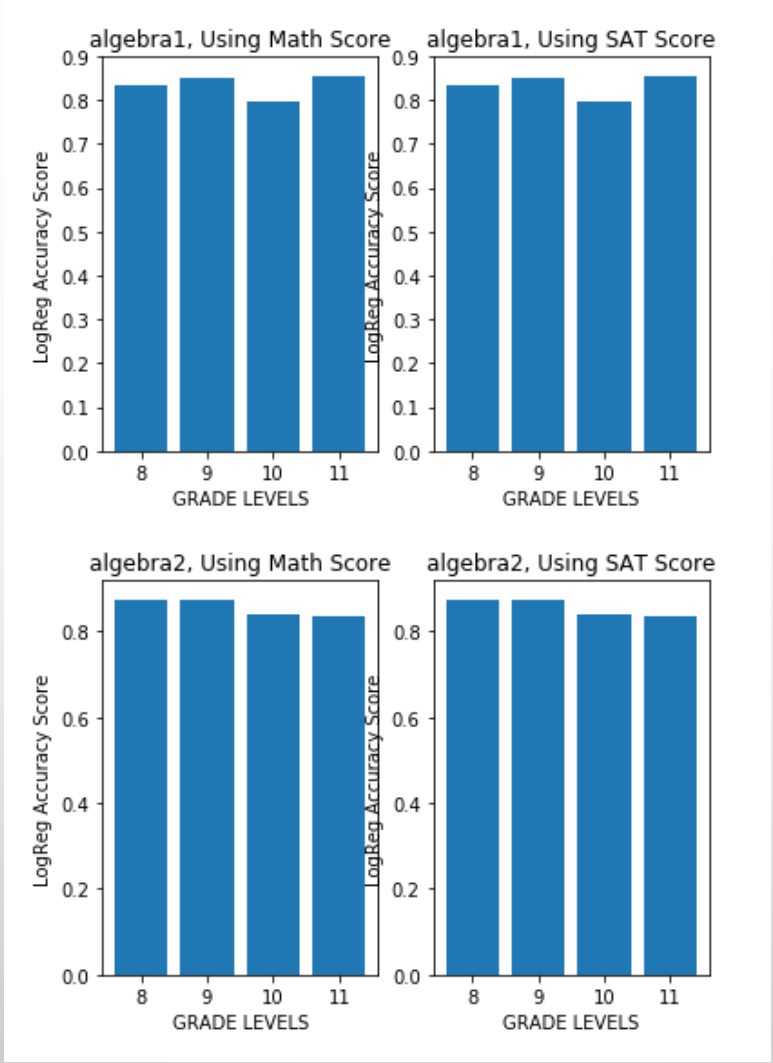
By simple algebraic manipulation, the probability that $Y = 1$ is

$$p = \frac{b^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{b^{\beta_0+\beta_1 x_1+\beta_2 x_2}+1} = \frac{1}{1+b^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}}.$$


Probability
0.0
0.2
0.4
0.6
0.8
1.0
linear
regression
line
0
500
1000
1500
2000
2500
below 0
Probability
0.0
0.2
0.4
0.6
0.8
1.0
logistic
regression
curve
0
500
1000
1500
2000
2500

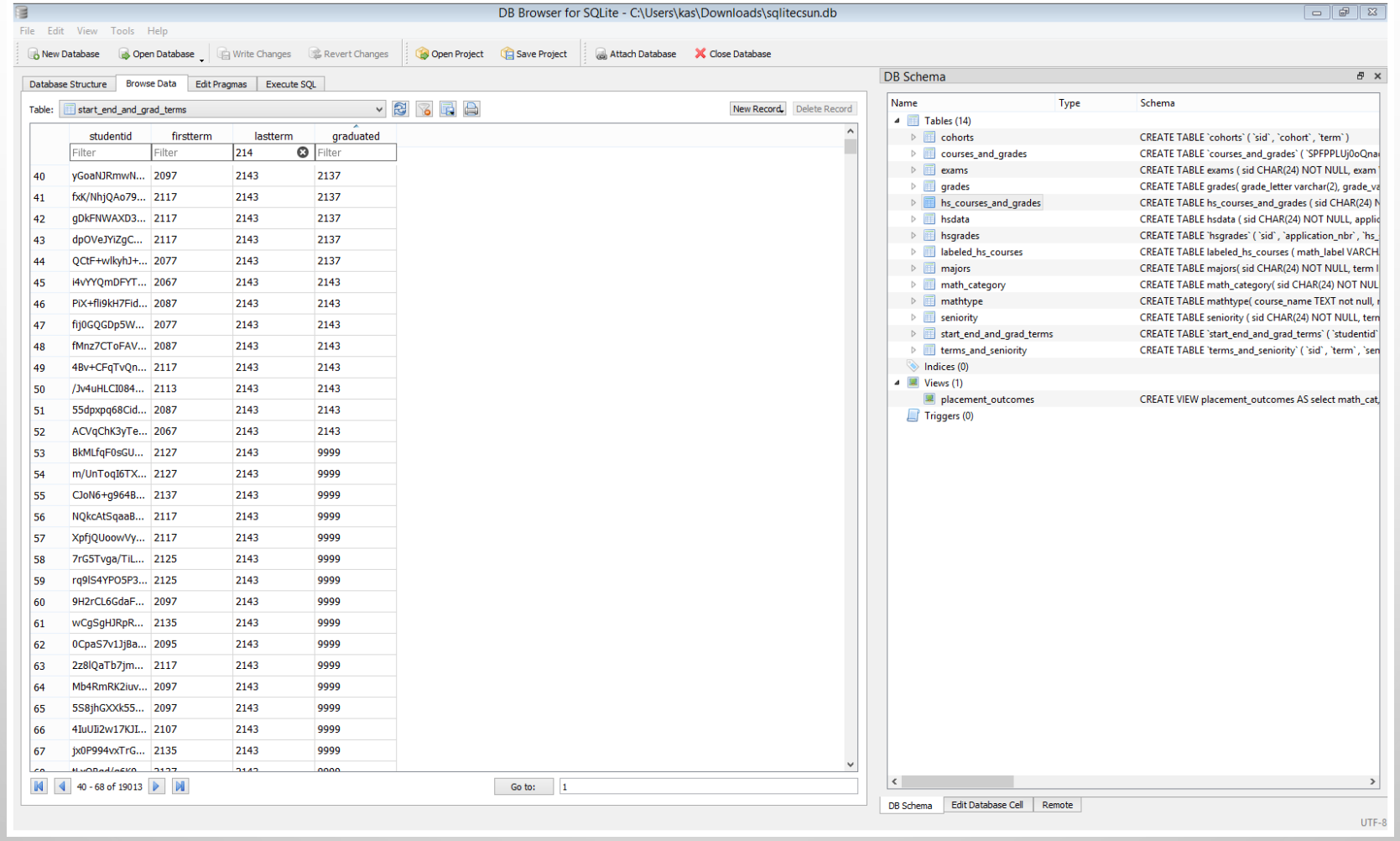
- SUBJECTS THAT WERE INCLUDED: ALGEBRA 1 AND 2, GEOMETRY AND TRIG
- PER HIGH SCHOOL MATH SUBJECT
- PER SCHOOL GRADE
- TWO SCORES FOR TWO MODELS: ONE BASED ON HS MATH, THE OTHER ON SAT SCORES
- COMPARE THE SCORES FROM BOTH MODELS

```
--------
ALGEBRA1
--------
Grade level 8:

Logistic regression results:
        sample size = 8358
        accuracy using math grade as predictor = 0.8522
        accuracy using SAT as predictor = 0.8498

Grade level 9:

Logistic regression results:
        sample size = 17376
        accuracy using math grade as predictor = 0.8520
        accuracy using SAT as predictor = 0.8527

Grade level 10:

Logistic regression results:
        sample size = 1879
        accuracy using math grade as predictor = 0.8468
        accuracy using SAT as predictor = 0.8170

Grade level 11:

Logistic regression results:
        sample size = 387
        accuracy using math grade as predictor = 0.8144
        accuracy using SAT as predictor = 0.9072



--------
ALGEBRA2
--------
Grade level 8:

Logistic regression results:
        sample size = 29
        accuracy using math grade as predictor = 0.5000
        accuracy using SAT as predictor = 1.0000
```


algebra1, Using Math Score
algebra1, Using SAT Score
algebra2, Using Math Score
algebra2, Using SAT Score
LogReg Accuracy Score
GRADE LEVELS
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
8
9
10
11

UNKNOWN FACTORS

- NO SOCIO-ECONOMIC DATA ON STUDENTS
  - SCHOOL QUALITY
  - PARENTS EDUCATION LEVEL
  - ACCESS TO TUTORING
- PART OF THE DATABASE CORRUPTED
  - COULD ONLY USE DATA UP TO 2014
- NO DATA ON MIDDLE SCHOOL


DB Browser for SQLite - C:\Users\kas\Downloads\sqlitecsun.db
File Edit View Tools Help
New Database
Open Database
Write Changes
Revert Changes
Open Project
Save Project
Attach Database
Close Database
Database Structure
Browse Data
Edit Pragmas
Execute SQL
Table: start_end_and_grad_terms
New Record
Delete Record
studentid firstterm lastterm graduated
Filter Filter 214 Filter
40 yGoaNJRmwN... 2097 2143 2137
41 fxK/NhjQAo79... 2117 2143 2137
42 gDkFNWAXD3... 2117 2143 2137
43 dpOVeJYiZgC... 2117 2143 2137
44 QCtF+wlkyhJ+... 2077 2143 2137
45 i4vYYQmDFYT... 2067 2143 2143
46 PiX+fli9kH7Fid... 2087 2143 2143
47 fij0GQGDp5W... 2077 2143 2143
48 fMnz7CToFAV... 2087 2143 2143
49 4Bv+CFqTvQn... 2117 2143 2143
50 /Jv4uHLCI084... 2113 2143 2143
51 55dpxpq68Cid... 2087 2143 2143
52 ACVqChK3yTe... 2067 2143 2143
53 BkMLfqF0sGU... 2127 2143 9999
54 m/UnToqI6TX... 2127 2143 9999
55 CJoN6+g964B... 2137 2143 9999
56 NQkcAtSqaaB... 2117 2143 9999
57 XpfjQUoowVy... 2117 2143 9999
58 7rG5Tvga/TiL... 2125 2143 9999
59 rq9lS4YPO5P3... 2125 2143 9999
60 9H2rCL6GdaF... 2097 2143 9999
61 wCgSgHJRpR... 2135 2143 9999
62 0CpaS7v1JjBa... 2095 2143 9999
63 2z8lQaTb7jm... 2117 2143 9999
64 Mb4RmRK2iuv... 2097 2143 9999
65 5S8jhGXXk55... 2097 2143 9999
66 4IuUIl2w17KJI... 2107 2143 9999
67 jx0P994vxTrG... 2135 2143 9999
40 - 68 of 19013
Go to: 1
DB Schema
Name Type Schema
Tables (14)
cohorts CREATE TABLE `cohorts` (`sid`, `cohort`, `term`)
courses_and_grades CREATE TABLE `courses_and_grades` (`SPFPPLUj0oQna
exams CREATE TABLE exams (sid CHAR(24) NOT NULL, exam
grades CREATE TABLE grades(grade_letter varchar(2), grade_va
hs_courses_and_grades CREATE TABLE hs_courses_and_grades (sid CHAR(24) N
hsdata CREATE TABLE hsdata (sid CHAR(24) NOT NULL, applic
hsgrades CREATE TABLE `hsgrades` (`sid`, `application_nbr`, `hs_
labeled_hs_courses CREATE TABLE labeled_hs_courses (math_label VARCH
majors CREATE TABLE majors(sid CHAR(24) NOT NULL, term
math_category CREATE TABLE math_category(sid CHAR(24) NOT NUL
mathtype CREATE TABLE mathtype(course_name TEXT not null,
seniority CREATE TABLE seniority (sid CHAR(24) NOT NULL, tern
start_end_and_grad_terms CREATE TABLE `start_end_and_grad_terms` (`studentid`
terms_and_seniority CREATE TABLE `terms_and_seniority` (`sid`, `term`, `sen
Indices (0)
Views (1)
placement_outcomes CREATE VIEW placement_outcomes AS select math_cat,
Triggers (0)
DB Schema
Edit Database Cell
Remote
UTF-8

CHALLENGES

- RETRIEVING DATA
- WORKING WITH DATA
- INTERPRETING THE RESULTS

RETRIEVING DATA

- SQL QUERY VS PANDAS
- MYSQL
- POSTGRES
- SQLITE

	sid	application_nbr	hs_subject	hs_crs_nbr	hs_grade_level	descr	fall_gr	spr_gr	sum_gr	honors
	Filter	Filter	Filter	Filter	Filter	tirg	Filter	Filter	Filter	Filter
1	Qt3NY94UMqZ...	00182043	2	6	11	Tirg/Math An...	C	B		
2	ubCYJc0dDMu...	00081101	2	8	11	Tirgonomorty...	C+	C		S

In [92]:

```
x = full.descr.unique().size
print("Starting number of unique course names: %s" %x)
full.descr = full.descr.str.lower()
full = full.replace(to_replace=r'ab.*$'       , value=''   , regex=True)
full = full.replace(to_replace=r'advanced'    , value=''   , regex=True)
full = full.replace(to_replace=r'.*trig.*$'   , value='trigonometry'  , regex=True)
full = full.replace(to_replace=r'.*geom.*$'   , value='geometry'  , regex=True)
full = full.replace(to_replace=r'.*stat.*$'   , value='statistics'  , regex=True)
full = full.replace(to_replace=r'a\|b$'       , value=''   , regex=True)
full = full.replace(to_replace=r'1/2.*$'      , value=' 2', regex=True)
full = full.replace(to_replace=r'honors '     , value=''   , regex=True)
full = full.replace(to_replace=r'^.*intro to ', value=''   , regex=True)
full = full.replace(to_replace=r'a/b/c.*$'    , value=''   , regex=True)
full = full.replace(to_replace=r'\bii.*$'     , value=' 2', regex=True)
full = full.replace(to_replace=r'\bi.*$'      , value=''   , regex=True)
full = full.replace(to_replace=r'\b1.*$'      , value=''   , regex=True)
full = full.replace(to_replace=r'  '          , value=' ' , regex=True)
full = full.replace(to_replace=r'\b$'         , value=''   , regex=True)
full.descr = full.descr.str.rstrip()
y = full.descr.unique().size
print("Final number of unique course names: %s" %y)
full.head()
```

```
Starting number of unique course names: 7003
Final number of unique course names: 2311
```