TOWARDS A STUDENTS' HIGH SCHOOL DROPOUT PREDICTION MODEL

A binary classification problem

Esther Dantra

PRIMARY GOALS

- Identify students at risk of dropping out before their 17th birthday
- Which factor has the largest impact on a student dropping out to help design intervention programmes

THE DATASET

Year: Left School	Student: Ethnicity Order	Student: Ethnic Group	Student: Student Gender	Student: Leaving Year Level	Student: Student Age	Student: Age (Retention to 17)	Qualification: Highest Attainment (5 groups)	Qualification: Level 1 or Above	Qualification: Level 2 or Above	Qualification: Level 3 or UE Award	School: School Type	School: School Sector	School: Authority	School: Definition	A
2021	7	Total	Male	Year 13	Age 20+	Stayed until age 17 or above	University Entrance	NCEA Level 1 or Above	NCEA Level 2 or Above	UE award or Level 3	Secondary (Year 9-15)	Secondary	State	Not Applicable	
2021	7	Total	Male	Year 13	Age 20+	Stayed until age 17 or above	University Entrance	NCEA Level 1 or Above	NCEA Level 2 or Above	UE award or Level 3	Secondary (Year 9-15)	Secondary	State	Designated Character School	
2021	7	Total	Male	Year 13	Age 20+	Stayed until age 17 or above	University Entrance	NCEA Level 1 or Above	NCEA Level 2 or Above	UE award or Level 3	Secondary (Year 9-15)	Secondary	State	School with Boarding Facilities	
2021	7	Total	Male	Year 13	Age 20+	Stayed until age 17 or above	University Entrance	NCEA Level 1 or Above	NCEA Level 2 or Above	UE award or Level 3	Secondary (Year 9-15)	Secondary	State	Not Applicable	
2021	7	Total	Male	Year 13	Age 20+	Stayed until age 17 or above	University Entrance	NCEA Level 1 or Above	NCEA Level 2 or Above	UE award or Level 3	Secondary (Year 9-15)	Secondary	State	School with Boarding Facilities	

```
ncea.shape
```

```
(347187, 30)
```

```
Year: Left School                                  int64
Student: Ethnicity Order                           int64
Student: Ethnic Group                             object
Student: Student Gender                           object
Student: Leaving Year Level                       object
Student: Student Age                               int64
Student: Age (Retention to 17)                    object
Qualification: Highest Attainment (5 groups)      object
Qualification: Level 1 or Above                   object
Qualification: Level 2 or Above                   object
Qualification: Level 3 or UE Award                object
School: School Type                               object
School: School Sector                             object
School: Authority                                 object
School: Definition                                object
School: Affiliation                               object
School: School Gender                             object
Region: General Electorate                        object
Region: Māori Electorate                          object
Region: TA Ward                                   object
Region: TA Board                                  object
Region: Education Area                            object
Region: Regional Council                          object
School: Decile                                    object
School: Quintile                                  object
Region: Territorial Authority                     object
Students (∑ Values)                                int64
dtype: object
```

FEATURES DESCRIPTION

Ethnic Group:
The ethnic group(s) the student identifies with, as recorded in ENROL. The ethnic group data is presented at level 1 and total response. That is, leavers are counted once in each ethnic group they identify with.
Students can be counted in up to three different ethnic groups but only once in the total, therefore the sum of the different ethnic groups will likely exceed the total. Ethnic groups should not be combined as some students will be counted twice.

School Type:
The type of the school, for example, Composite (Year 1 to 15), Secondary (Year 7 to 15), Secondary (Year 9 to 15) the leaver left from.

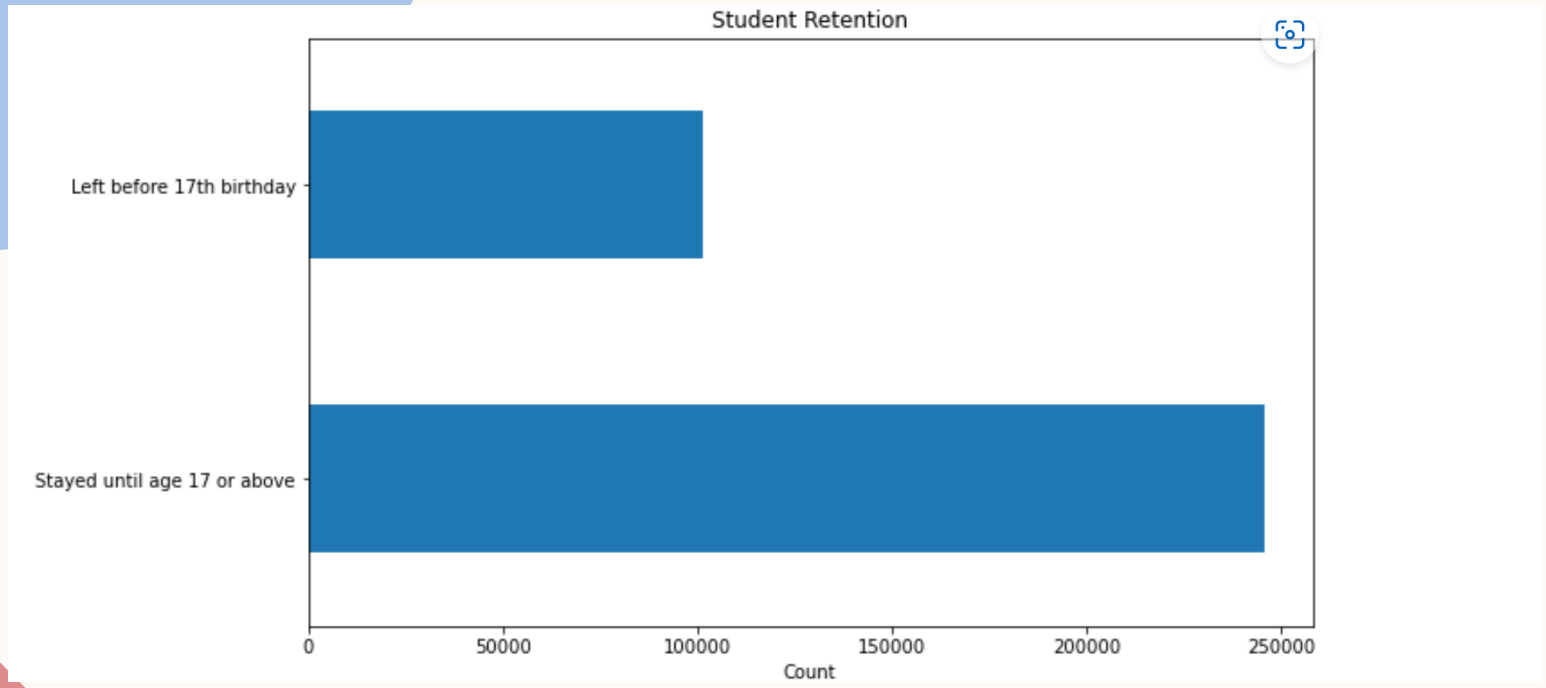
School Gender:
The gender of the students that a school caters for, for example, co-educational, or single sex.

Decile:
The decile assigned to the school. Students from low socio-economic communities face more barriers to learning than students from high socio-economic communities. Schools that draw their roll from these low socio-economic communities are given greater funding to combat these barriers. The mechanism used to calculate and allocate this additional funding is most often known as school deciles.

NCEA level 1 qualification:
Below Level 1 or gained level 1

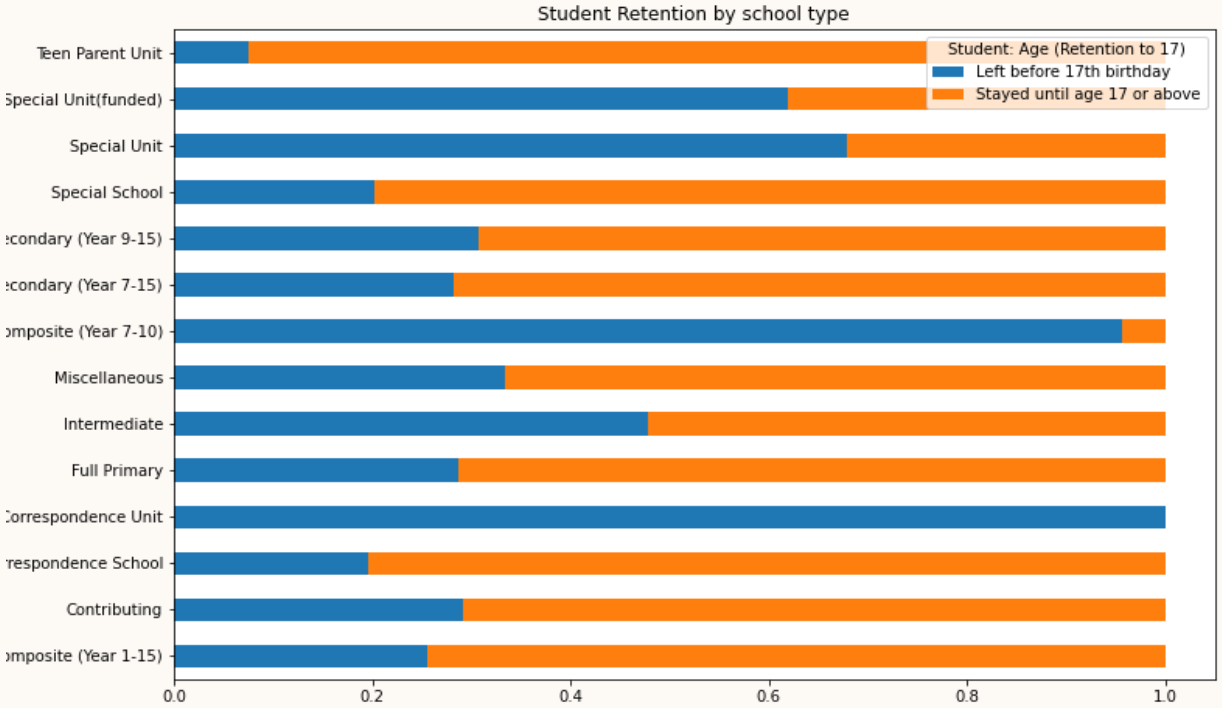
Student gender:
Male or Female at enrol

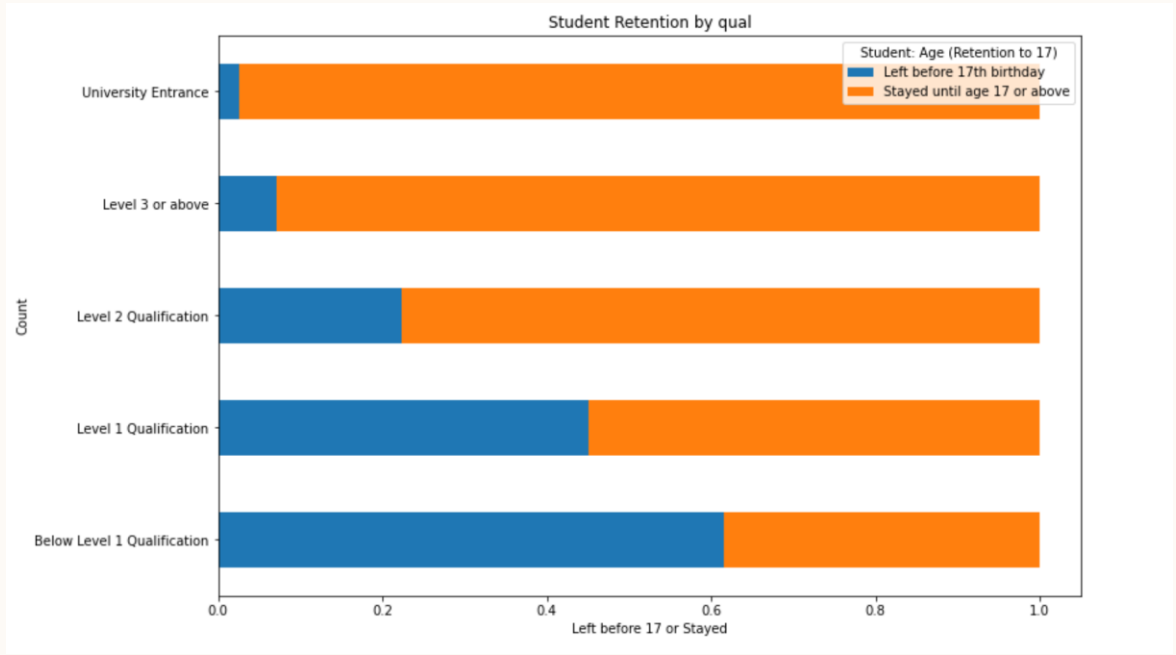
INITIAL INSIGHTS

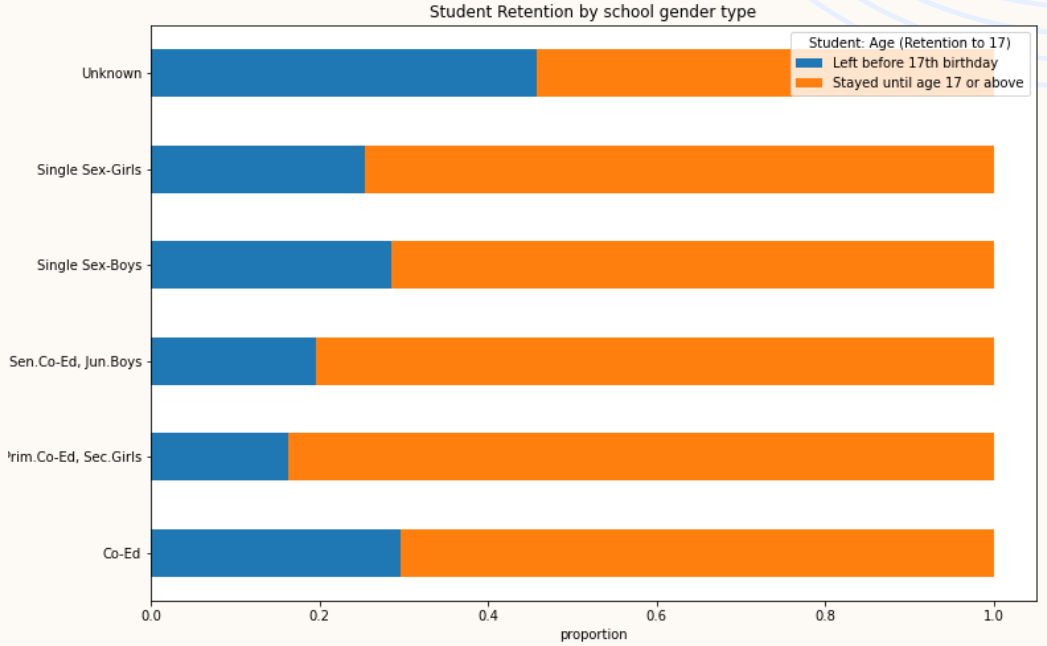


- 100 000 datapoints for students who have dropped out
- No need for over sampling

FURTHER INSIGHTS







FEATURES CHOSEN

	Ethnicity	Gender	SchoolType	SchoolGender	Level1_qual	Region	Decile	QualProportion
0	1	Female	Secondary (Year 7-15)	Co-Ed	Below NCEA Level 1	Whangarei	4.0	0.382270
1	1	Female	Secondary (Year 7-15)	Co-Ed	Below NCEA Level 1	Napier	2.0	0.382270
2	1	Female	Secondary (Year 7-15)	Co-Ed	Below NCEA Level 1	Napier	2.0	0.382270
3	1	Female	Secondary (Year 7-15)	Co-Ed	NCEA Level 1 or Above	Tukituki	1.0	0.382270
4	1	Female	Secondary (Year 7-15)	Co-Ed	NCEA Level 1 or Above	Botany	1.0	0.382270
...	...	...	...	...	...	...	...	...

- Features which will affect the model were removed – eg: passed ncea level 2 or 3 because this happens after 17th birthday
- Many columns were doubling on information, city as well as district. Chose district over city. District is more detailed information about a student's region than city)

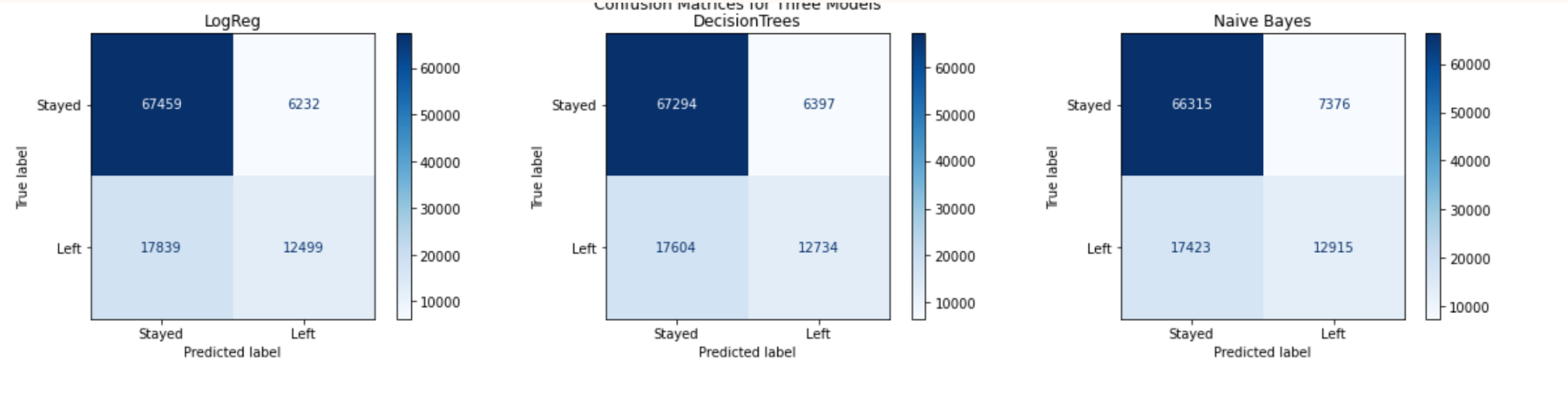
MODEL - COMPARISON

Logistic Regression	
Precision:	0.67
F1 Score:	0.51
Recall:	0.41
Accuracy:	0.77

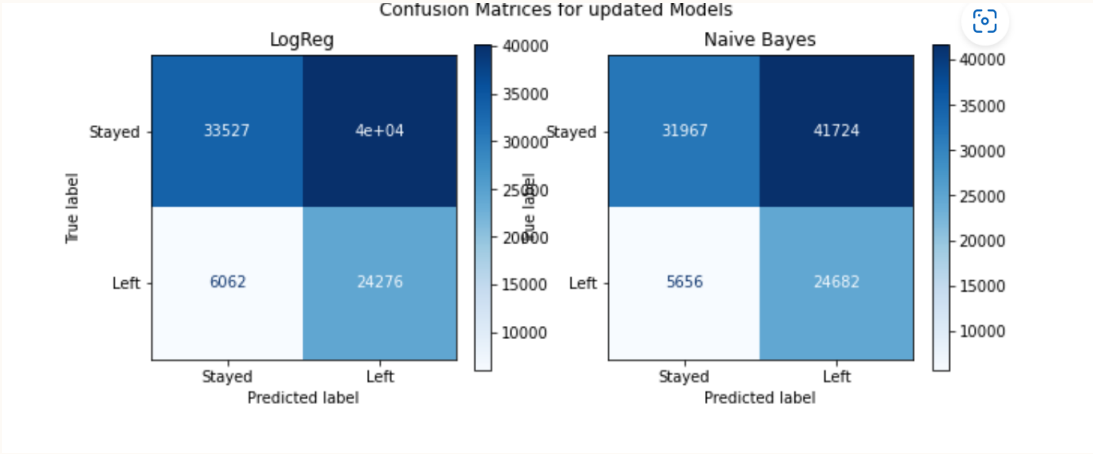
Decision Tree	
Precision:	0.67
F1 Score:	0.48
Recall:	0.38
Accuracy:	0.76

Naïve Bayes	
Precision:	0.35
F1 Score:	0.47
Recall:	0.69
Accuracy:	0.54

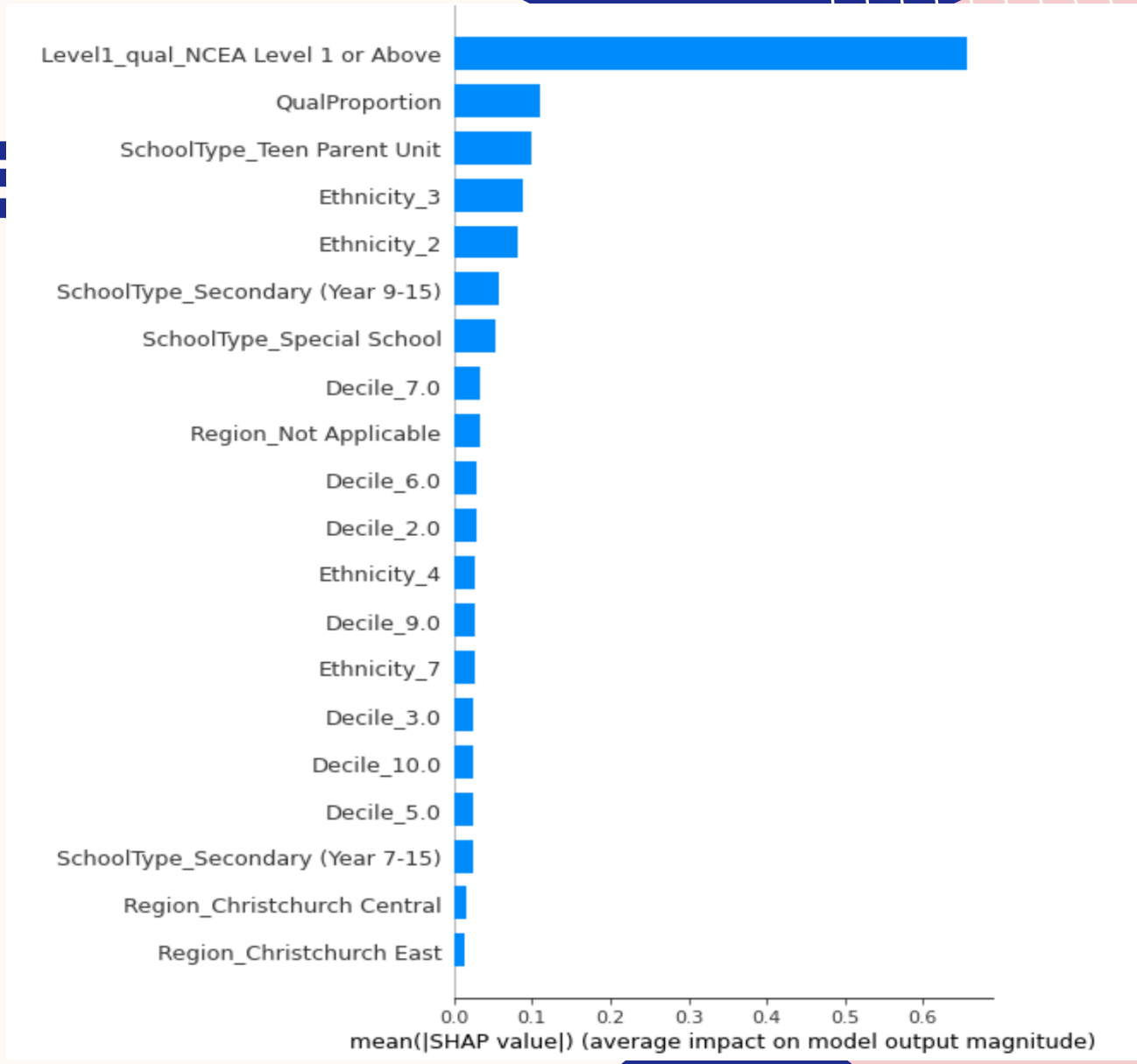
MODEL COMPARISON



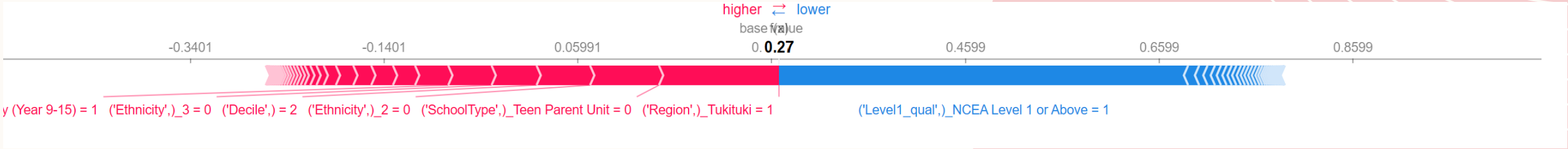
After changing thresholds:

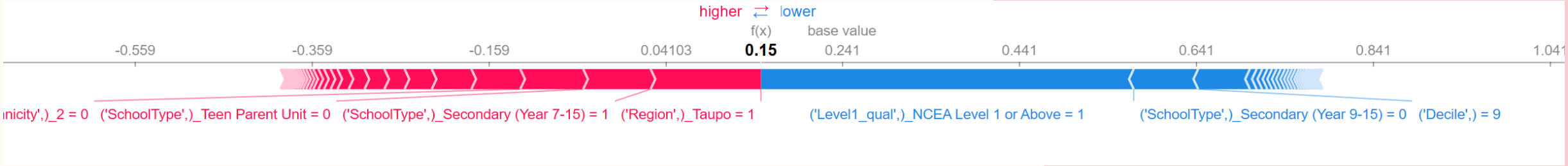


FEATURE IMPORTANCE



FEATURE IMPORTANCE





- Confirms that passing level 1 is the key feature
- Both of these students have stayed at school, probability is closer to 0 than 1.

SUMMARY

- Logistic regression is best so far
- Passing Level 1 NCEA is a big indicator
- Ethnicity, school type are also big indicators
- Data on unexplained absences, classroom engagement etc is required to get better accuracy