Predicting Students at Risk of Becoming NEET (Not in Education, Employment or Training): Methodology Documentation

Submitted By

Organisation: Data Science for Social Good (DSSG) United Kingdom 2022

Contributors: Abhijeet Mulgund, Pranjusmrita Kalita, Rachel Humphries, Vanshika Namdev

Technical Mentor: Mihir Mehta

Project Manager: Satyam Bhagwanani

Table of Contents

Overview

Data Science for Social Good (DSSG)

DSSG is a 12-week summer programme providing not-for-profit organisations and government bodies unprecedented access to inspiring, top-tier data science talent.

DSSG helps partners to achieve more with their data by enhancing their services, interventions, and outreach, helping fulfil their mission of improving the world and people's lives. Project teams build their capacity to use cutting-edge quantitative methods to address societal challenges in areas such as education, health, energy, public safety, transportation, and economic development.

For more information about the programme, see:
https://warwick.ac.uk/research/data-science/warwick-data/dssgx/

Project Title

Predicting Students at Risk of Becoming NEET (not in education, employment, or training).

Partner Organisation Name & Background

Buckinghamshire Council – a unitary local authority in South-East England responsible for providing all local government services in the region and serving a population of approximately 550k people.

Problem Statement

Between 2018 and 2020, Buckinghamshire county had a NEET rate of above 2% and Unknown destination rate of above 5% for young people aged 17 to 18.

Studies have shown that time spent NEET can have a detrimental effect on physical and mental health, increasing the likelihood of unemployment, low wages, or low quality of work later on in life. Buckinghamshire Council wanted to identify students' risk of becoming NEET in years 12 or 13 (ages 17-18), by the end of years 10 or 11 (ages 14-16) so that they could target the right pupils with early intervention programmes. It is hoped that doing so will improve the life chances of those young people who receive intervention that they otherwise may not have done.

Project Objective

The objective of this project was to predict which pupils are at risk of becoming NEET in the future so that the council and schools can ensure they are targeting the right pupils with early intervention programmes.

Solution

The project team have built several artefacts to support the council in this objective.

A predictive model that:

- Predicts the risk of becoming NEET for 26,592 students currently in school years 7 to 10
- Identifies students' key risk factors contributing to their probability of becoming NEET
- Identifies if students with an UNKNOWN status are likely to become NEET

A PowerBI dashboard that:

- Allows Buckinghamshire Council to view the insights generated by the model for each student
- Provides insights about schools and school areas with a higher rate of NEET
- Allows the council to view insights about a larger cohort of 61,761 unique students from the NCCIS dataset (2017-2022), in school years 6 to 13

Results & Value Added

Overall, the predictive model achieved an accuracy of 92.8% and an F2-score of 47.8%.

The accuracy score indicates that the model is able to correctly predict the outcome of a student (NEET or not NEET) 92.8% of the time.

The F2-score is a measure of how well the model minimises false positives (students predicted as high risk who are unlikely to become NEET) and false negatives (students who are likely to become NEET but are not predicted as high risk). The F2-score was chosen over the F1-score to measure the performance of the model since the latter gives equal weight to false positives and false negatives, while the former gives a higher weight to minimising false negatives, which was considered more important by the council.

At a UK level the tool has the potential to identify 22% (4,193) more students per year who become NEET as compared to the already existing Risk of NEET Indicator (RONI) tool, which is used by some local authorities in the UK. When tested on the same dataset, the RONI tool achieved an accuracy of 85.5%.

It also flags 51% fewer students who never became NEET as compared to RONI, and therefore has the potential to save significant operational costs and resources for councils across the country.

Next Steps

The tool has been designed with the aim of being adopted by Buckinghamshire Council as well as other Councils and Local authorities to ensure that students who are identified as being at high risk of becoming NEET are able to benefit from early intervention programmes, thereby preventing NEET outcomes and contributing to a better long-term quality of life.

While the model does outperform existing tools such as RONI, it is anticipated that it could be further improved by including datasets and features that are known to be predictive of poor outcomes for young people. Specifically, the council is encouraged to integrate data from sources such as Early Help & Social Care, Revenues & Benefits, and from wider public services such as the NHS and the police to further improve performance. A detailed list of possible datasets is provided in the section on Suggestions for Further Work.

Methodology Documentation

This document provides a brief overview of the methodology the team followed to produce the above outputs. The purpose of this document is to support other councils to be able to reproduce the analysis and modelling using their own data. It should be read in conjunction with the code and README that can be found on the project GitHub page: https://github.com/DSSGxUK/s22_buckinghamshire.

This documentation covers:

- Data Schema
- Data & Modelling Pipeline and Approach
- Outputs
- Suggestions for Further Work

Data Schema

The following datasets were provided by Buckinghamshire council and used by the team to carry out the modelling and analysis. Most of the datasets follow schemas set by central government, and where available links have been provided which describe the metadata and data fields in detail.

Attendance Dataset - Provides data on the attendance of students along with features like termly sessions, absences, and reasons for absences (exclusions, late entry, etc) – Attendance_Schema.

School Census Dataset - Provides demographic information about students, for example: Gender, Ethnicity, Age and Language, as well as other features such as whether the student receives Free School Meals (FSM) or has Special Educational Needs (SEN) - Census_Schema.

NCCIS Dataset - This dataset holds information required by Local Authorities to support young people to engage in education and training. Some of the important variables captured in the dataset are student's characteristic codes (e.g. if they are a carer, pregnant etc), activity codes, Special Educational Needs, and their level of need. The final outcome variable of whether a student is NEET or UNKNOWN is extracted from this dataset - CCIS_Schema.

KS4 Dataset - Provides information related to student's grades, eligibility for Free School Meals and Income deprivation index - KS4_Schema.

School Information Dataset - Provides details on school areas, postcodes, and school names.

The table below shows the years covered by each of the first four datasets in the data that was provided by the partner:

	2015	2016	2017	2018	2019	2020	2021	2022
Attendance	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
School Census			🟨	🟨	🟨	🟨	🟨	🟨
NCCIS		🟩	🟩	🟩	🟩	🟩	🟩	🟩
KS4	🟧	🟧	🟧	🟧	🟧			

Data & Modelling Pipeline and Approach

Two pipelines were developed for the project, to prepare the data, and to manage the modelling process. A Python package called Data Version Control (https://dvc.org/doc/api-reference) was used to develop the modelling pipeline, which provides a simple mechanism to reproduce the project outputs using different datasets. Further details of how to run the pipeline can be found in the README on the project GitHub page: https://github.com/DSSGxUK/s22_buckinghamshire.

The diagram below provides a simple overview of the end-to-end pipeline:

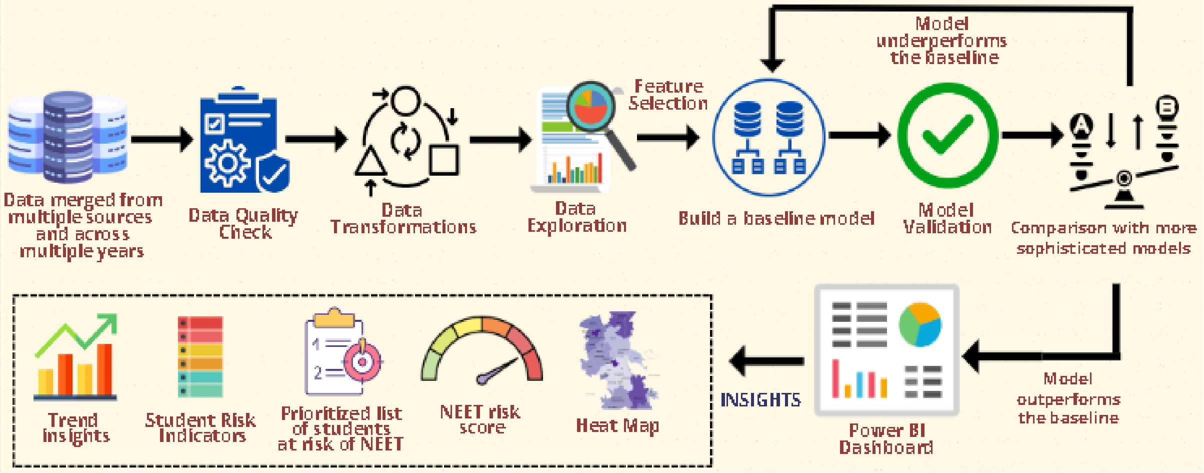

Data merged from multiple sources and across multiple years
Data Quality Check
Data Transformations
Data Exploration
Feature Selection
Build a baseline model
Model underperforms the baseline
Model Validation
Comparison with more sophisticated models
Model outperforms the baseline
Power BI Dashboard
INSIGHTS
Trend Insights
Student Risk Indicators
Prioritized list of students at risk of NEET
NEET risk score
Heat Map

Fig 1.1: End-to-End pipeline of data preparation and modelling steps

Below is a brief overview of the key stages in the workflow.

Generate Datasets for Modelling

At this step, the pipeline:

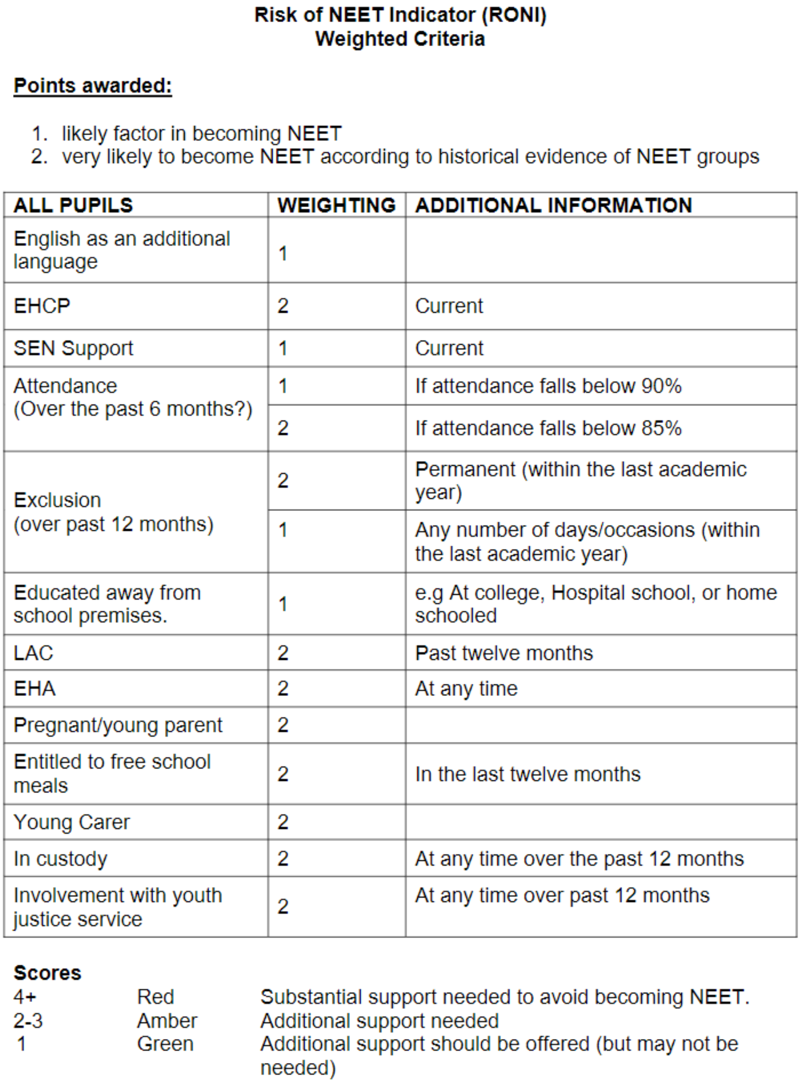
- Merges each dataset (e.g. Census, Attendance, etc) across all the years
- Splits categorical variables into binary columns containing 0's or 1's
- Drops columns which aren't required for modelling or for which there wasn't data available before Year 11
- Renames columns
- Removes columns with more than 60% missing values and removes any empty rows
- Replaces columns that have string values of ‘Y’ or ‘N’ with 0 or 1
- Outputs two datasets ready for modelling:
    - A single UPN dataset, with one row per student
    - A multiple UPN dataset, where a student can have multiple observations across different years
- Splits the data into training (80%) and testing (20%) sets

Since the dataset was highly imbalanced (only 1 to 2% of students had a NEET outcome), more samples were required to train the model. As such, Random Oversampling was used, setting the *sampling_strategy* attribute equal to 0.1 and *shrinkage* attribute equal to 0.01.

The reason for outputting two datasets was to test which one would achieve better performance. Specifically, we were interested to understand whether having the same child appear multiple times in the dataset may help the model to understand how change in a student's circumstances may affect their risk of becoming NEET. Ultimately, the single UPN dataset proved to be better for modelling and all the reported results pertain to that dataset.

Baseline Model – Risk of NEET Indicator (RONI) Tool

The RONI Tool was developed by the National Foundation for Educational Research in 2012 (see: https://www.nfer.ac.uk/publications/INDI02/INDI02.pdf). It assists secondary schools to identify students at the risk of becoming NEET once they leave compulsory education. It uses a simple rule-based approach that assigns risk scores to students based on various risk factors.



Risk of NEET Indicator (RONI)
Weighted Criteria

Points awarded:

1. likely factor in becoming NEET
2. very likely to become NEET according to historical evidence of NEET groups

ALL PUPILS	WEIGHTING	ADDITIONAL INFORMATION
English as an additional language	1	
EHCP	2	Current
SEN Support	1	Current
Attendance (Over the past 6 months?)	1	If attendance falls below 90%
	2	If attendance falls below 85%
Exclusion (over past 12 months)	2	Permanent (within the last academic year)
	1	Any number of days/occasions (within the last academic year)
Educated away from school premises.	1	e.g At college, Hospital school, or home schooled
LAC	2	Past twelve months
EHA	2	At any time
Pregnant/young parent	2	
Entitled to free school meals	2	In the last twelve months
Young Carer	2	
In custody	2	At any time over the past 12 months
Involvement with youth justice service	2	At any time over past 12 months

Scores

4+	Red	Substantial support needed to avoid becoming NEET.
2-3	Amber	Additional support needed
1	Green	Additional support should be offered (but may not be needed)

Fig 1.2: Risk factors and weightings used by the RONI tool

Before developing a predictive model using machine learning, we assessed our dataset against the RONI criteria to produce a baseline model. When we ran the

RONI tool on our test data, this achieved an accuracy of 85.5% and an F2 score of 0.32.

Model Training and Cross-Validation

The next stage in the pipeline involves the training the model, applying cross-validation and searching for the best hyperparameters. At this step, the pipeline:

- Applies k-fold cross-validation
- Uses the GridSearchCV method from scikit-learn to search for the best model parameters

Model Evaluation

The next stage in the pipeline involves evaluating the model. At this step, the pipeline:

- Evaluates the RONI tool's performance
- Retrains the model with new incoming data and saves the model with the best threshold
- Applies the chosen model on the test data to output a final model performance score

Using the data provided by Buckinghamshire, the best performing model was Gradient Boosted Trees from the LightGBM package in Python, which achieved an accuracy of 92.8% and F2 score of 47.8%.

Generate Datasets for Predictions and Final Output

The final stage in the pipeline is to generate the final predictions and output datasets. At this step, the pipeline:

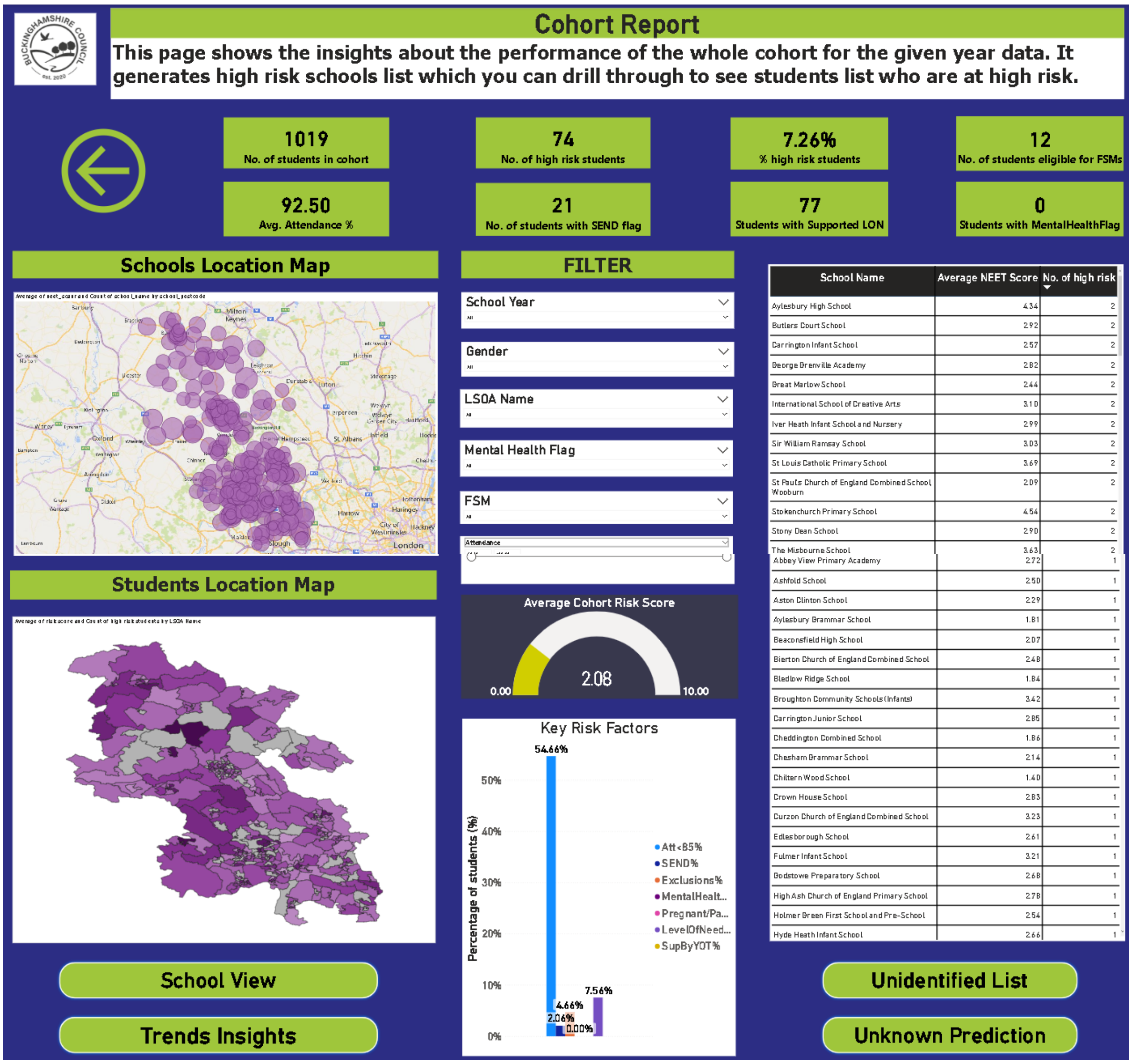
- Creates datasets with current Year 7-10 students and students with unknown destinations to predict on
- Executes the model on unseen data and generates final predictions in the form of a CSV file
- Generates feature importance using SHAP values
- Returns the RONI score
- Returns scaled probability scores for a student at risk of becoming NEET (between 1-10)

Based on Buckinghamshire’s data, the features that were most predictive in assessing NEET risk were:

- Level of Need: Supported
- Total Absences
- Characteristic: Mental Health
- Unauthorised Absences (Other)
- Characteristic: Parent
- Authorised Absences (Exclusion)

Outputs

The following are some snippets from the Power BI dashboard delivered to Buckinghamshire for viewing future predictions of NEET, key risk factors contributing to their risk, and trends in the data. It is important to note that the dashboard has been loaded with fake data but has the original schema.



Cohort Report

This page shows the insights about the performance of the whole cohort for the given year data. It generates high risk schools list which you can drill through to see students list who are at high risk.

1019 No. of students in cohort	74 No. of high risk students	7.26% % high risk students	12 No. of students eligible for FSMs
92.50 Avg. Attendance %	21 No. of students with SEND flag	77 Students with Supported LON	0 Students with MentalHealthFlag

FILTER

School Year: All

Gender: All

LSOA Name: All

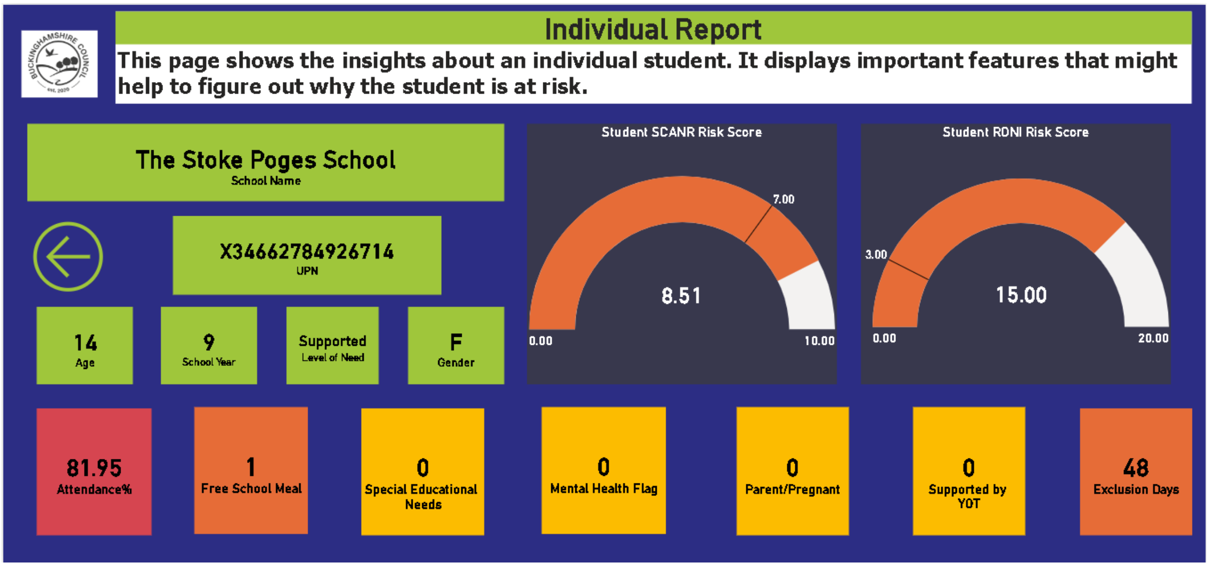
Mental Health Flag: All

FSM: All

Attendance

School Name	Average NEET Score	No. of high risk
Aylesbury High School	4.34	2
Butlers Court School	2.92	2
Carrington Infant School	2.57	2
George Grenville Academy	2.82	2
Great Marlow School	2.44	2
International School of Creative Arts	3.10	2
Iver Heath Infant School and Nursery	2.99	2
Sir William Ramsay School	3.03	2
St Louis Catholic Primary School	3.69	2
St Pauls Church of England Combined School, Wooburn	2.09	2
Stokenchurch Primary School	4.54	2
Stony Dean School	2.90	2
The Misbourne School	3.63	2
Abbey View Primary Academy	2.72	1
Ashfold School	2.50	1
Aston Clinton School	2.29	1
Aylesbury Grammar School	1.81	1
Beaconsfield High School	2.07	1
Bierton Church of England Combined School	2.48	1
Bledlow Ridge School	1.84	1
Broughton Community Schools (Infants)	3.42	1
Carrington Junior School	2.85	1
Cheddington Combined School	1.86	1
Chesham Grammar School	2.14	1
Chiltern Wood School	1.40	1
Crown House School	2.83	1
Curzon Church of England Combined School	3.23	1
Edlesborough School	2.61	1
Fulmer Infant School	3.21	1
Godstowe Preparatory School	2.68	1
High Ash Church of England Primary School	2.78	1
Holmer Green First School and Pre-School	2.54	1
Hyde Heath Infant School	2.66	1

Fig 2.1: Dashboard page with NEET predictions for the complete cohort in Buckinghamshire

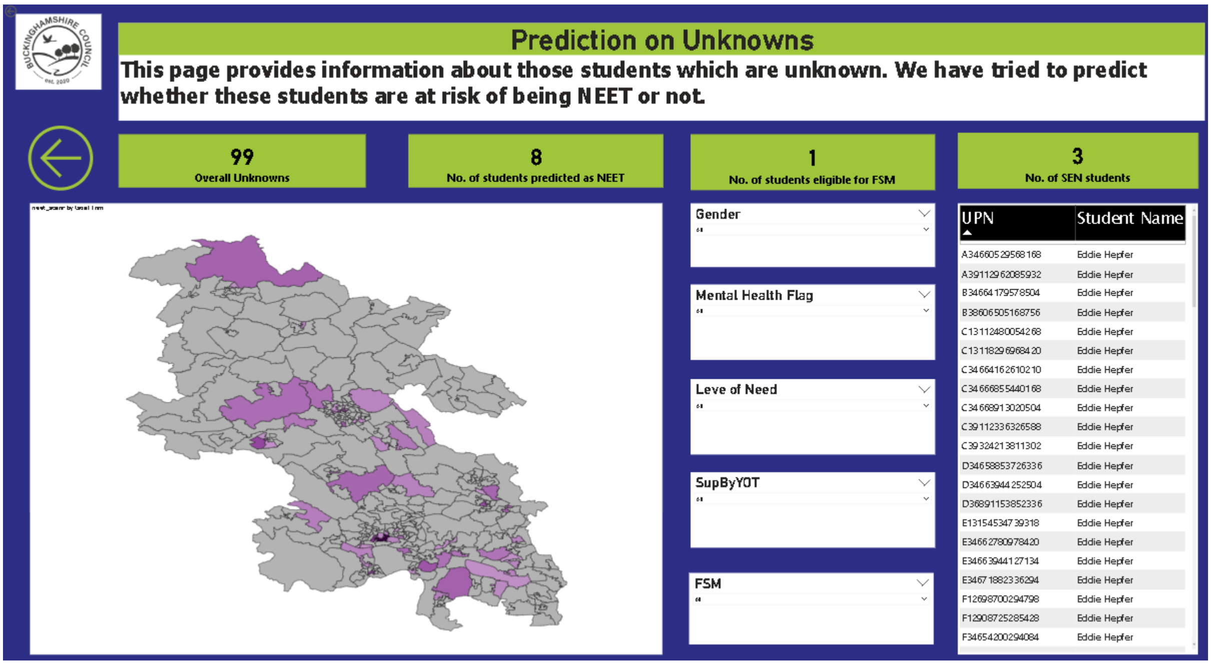


Individual Report

This page shows the insights about an individual student. It displays important features that might help to figure out why the student is at risk.

School Name	The Stoke Poges School
UPN	X34662784926714
Age	14
School Year	9
Level of Need	Supported
Gender	F
Student SCANR Risk Score	8.51
Student RONI Risk Score	15.00
Attendance%	81.95
Free School Meal	1
Special Educational Needs	0
Mental Health Flag	0
Parent/Pregnant	0
Supported by YOT	0
Exclusion Days	48

Fig 2.2: Dashboard page showing information related to an individual student



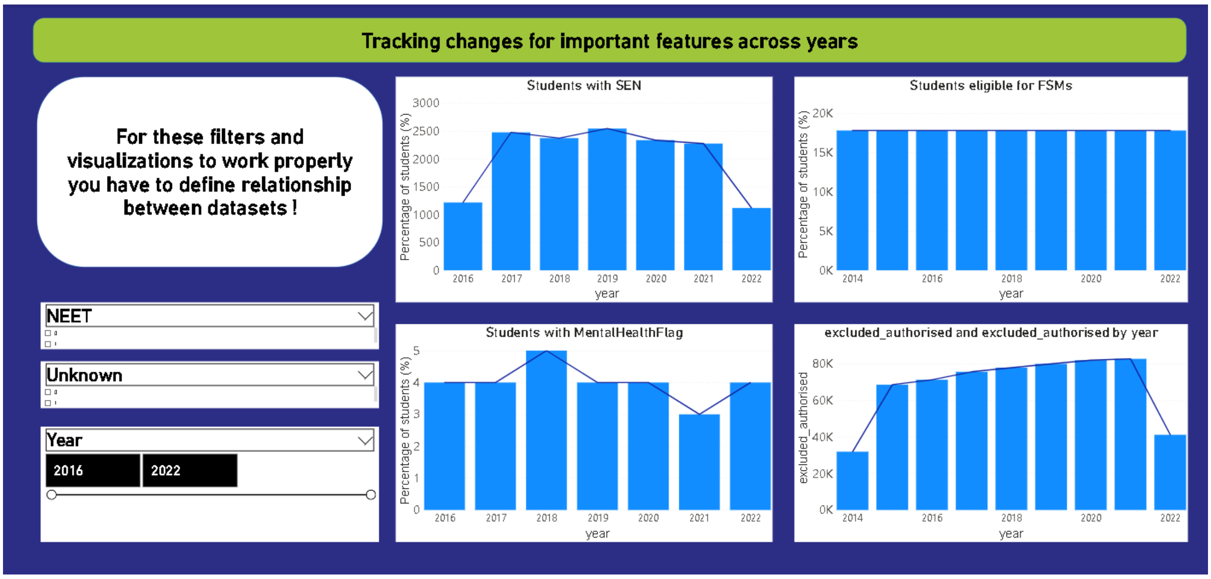
Prediction on Unknowns

This page provides information about those students which are unknown. We have tried to predict whether these students are at risk of being NEET or not.

Overall Unknowns	99
No. of students predicted as NEET	8
No. of students eligible for FSM	1
No. of SEN students	3

UPN	Student Name
A346605 29568 168	Eddie Hepfer
A39112962085932	Eddie Hepfer
B34664179578504	Eddie Hepfer
B38606505168756	Eddie Hepfer
C13112480054268	Eddie Hepfer
C13118296968420	Eddie Hepfer
C34664162610210	Eddie Hepfer
C34666855440168	Eddie Hepfer
C34668913020504	Eddie Hepfer
C39112336326588	Eddie Hepfer
C39324213811302	Eddie Hepfer
D34658853726336	Eddie Hepfer
D34663944252504	Eddie Hepfer
D36891153852336	Eddie Hepfer
E13154534739318	Eddie Hepfer
E34662780978420	Eddie Hepfer
E34663944127134	Eddie Hepfer
E34671882336294	Eddie Hepfer
F12698700294798	Eddie Hepfer
F12908725285428	Eddie Hepfer
F34654200294084	Eddie Hepfer

Fig 2.3: Dashboard page with predictions on UNKNOWN students



Tracking changes for important features across years

For these filters and visualizations to work properly you have to define relationship between datasets !

Fig 1.4: Dashboard page for viewing trends in historical data

Suggestions for Further Work

While the model does outperform existing tools such as RONI, it is anticipated that it could be further improved by including datasets and features that are known to be predictive of poor outcomes for young people. Specifically, councils are encouraged to integrate data from the following sources to see whether it would improve model performance:

Social Care & Early Help – data on whether a child is known to Children's Social Care or Early Help services. This may include datasets or fields relating to the following:

- Children's Social Care Contacts
- Children's Social Care Assessments
- Early Help
- Children in Need
- Child Protection
- Looked After Children
- Children with Disabilities
- Children Missing from Education
- Child Sexual Exploitation
- Domestic Violence Incidents
- Troubled Families
- Young Carers
- Youth Offending Services

Revenues & Benefits – specifically data relating to debt and benefits. This may include datasets or fields relating to the following:

- Council Tax Debt / Arrears
- Council Tax Reduction

- Housing Benefit
- Household Composition

In addition to the above, if datasets from other public services could also be added, for example the NHS or local police force, this has even more potential to improve the model. Bristol City Council are one authority who have successfully worked across the public services system to link data to improve outcomes for young people, including building a predictive model for NEETs. Details of their work can be found here: https://www.bristol.gov.uk/council-and-mayor/policies-plans-and-strategies/social-care-and-health/insight-bristol-and-the-think-family-database.