Predicting Education Performance at Local Authority Level

Lorenzo Ognibeni, Rhiana Dennis-Small, Harinee Devarajan

Working Paper No. 003 August 2025

EMAP Working Paper Series

Queen Mary University of London

Predicting Education Performance at Local Authority Level

Group B: Lorenzo Ognibeni, Rhiana Dennis-Small, Harinee Devarajan

Non-technical Summary

This report will explore the drivers behind education attainment scores at Local Authority (LA) level. This aim is to be able to support policymakers to be able to improve education in line with the current Levelling Up agenda. Improving attainment has many benefits, such as improving the average standard of an individual's skill set which can lead to higher productivity and longer-term economic growth. This is particularly important given the international competition the country faces.

Attainment 8 scores are a standard measure of how well pupils do at GCSE across eight subjects, with a higher weighting given to Maths and English GCSE scores. We measure performance using the average Attainment 8 score of all pupils attending state-funded schools in a LA. As we have used a continuous variable in Attainment 8 scores as our target, we have a supervised regression problem. A supervised regression problem is where we aim to predict a target variable with a given set of predictor variables.

We source Local Authority level data publicly available from the Department for Education which includes a range of information about schools such as their finances, their teacher workforce, and pupil characteristics. We compile this to form a dataset from 2015-2021 which resulted in 826 observations and 533 features in total.

For the analysis we focus on supervised machine learning techniques as those are the methods for making predictions. Firstly, we use tree-based methods that provide a good prediction accuracy but overfitted our model. This means although the bias might be low, the variance is relatively high which is a constant trade-off faced in machine learning. We therefore also use an unsupervised machine learning technique to simplify our dataset. Unsupervised machine learning techniques aim to understand the variables rather than predicting a target variable. This ends up not being as effective as we hoped in improving the model for any of the tree-based methods that we use. We decide to stretch our model by exploring the use of linear methods, specifically comparing across the different penalised regression methods. This gives us the best prediction accuracy across all the previous methods that we have used.

Using the results from this final method, we identify that the most important variables for estimating attainment could be split into four distinct categories: inequalities, absences from school, targeted funding, and teacher quality. We suggest potential policies that we believe will make an impact specific to these categories. However, we find that the specific variables we identify are either not in direct control of education policymakers, or there needs to be further research to identify the factors that affect these variables. Overall, our results help policymakers focus on areas that can make a key difference to improve attainment and lead to improved societal outcomes.

Introduction

The General Certificate of Secondary Education (GCSE) is an exam taken by students (usually 15- to 16-year-olds) in England, Wales, and Northern Ireland. GCSE results are a key measure of educational attainment and academic achievement, serving as an important benchmark for assessing the performance of not just students but schools too. Encouragingly, education levels in the UK have increased in the last 20 years, with the share of students achieving 5 good GCSES going from under 40% in the early 1990s, increasing to 82% in 2012 (Institute for Fiscal Studies, 2022). Nonetheless, improving pupil attainment is a constant area of policy focus, given the enormous benefits associated with higher human capital accumulation, which can lead to higher productivity and economic growth at a societal level in the long term. Enhancing educational attainment also has significant benefits to the student beyond improved labour market outcomes, in the form of greater personal fulfilment and wellbeing as education is about personal development as well as academic achievement.

Our project aims to explore educational attainment at Local Authority (LA) level. Attainment 8 scores are a standard measure of how well pupils do at GCSE across 8 subjects, with a higher weighting given to Maths and English GCSE scores. We measure performance using the average Attainment 8 score of all pupils attending state-funded schools in a LA. We aim to understand the factors at LA level which are important for predicting Attainment 8 scores. This is important for understanding how best to bridge disparities between LAs in terms of educational outcomes by identifying the underlying factors which need to be addressed. This is another important consideration on the policy agenda given the focus on Levelling Up.

We use existing studies to help us identify potential factors to consider when analysing attainment. For example, Baker et al. (2002) finds that within-school differences are larger than between-school differences. In other words, the influence that a school has on a student's attainment is large, even if other characteristics such as family aspects are dominant. In addition, a paper by the OECD (2019) finds that students tend to do worse when their peers are also low achievers or socially disadvantaged. Therefore, we think looking at the characteristics of the schools and students is important to understand differences in attainment.

Given we are interested in making predictions, we have focused primarily on supervised machine learning techniques. These are methods that predict the value of a target variable using a given set of predictor variables. Given that we use a continuous variable in Attainment 8 scores as our target, we are dealing with a supervised regression problem.

Data

Target

We have the average Attainment 8 score of all pupils that attend a state-funded school in a Local Authority. Figure 1 shows that within each year, there is a widespread in performance between Local Authorities in terms of GCSE results. Figure 2 shows that the average score varies between 46 and 51 depending on the year, bearing in mind that the theoretical maximum Attainment 8 score is 90. There was a spike in 2020 because exams were cancelled due to COVID, and teacher-predicted grades were used instead of exam grades, leading to widespread grade inflation.

Figure 1:
Density plot of Average Attainment 8 scores per year.

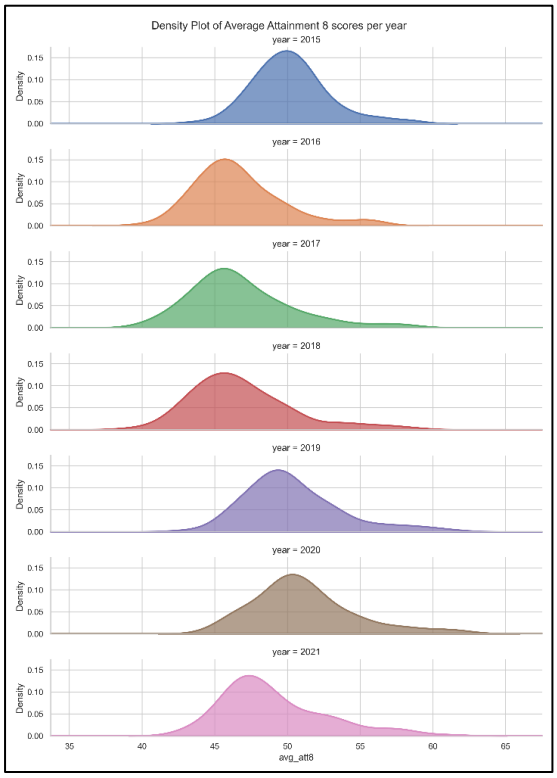
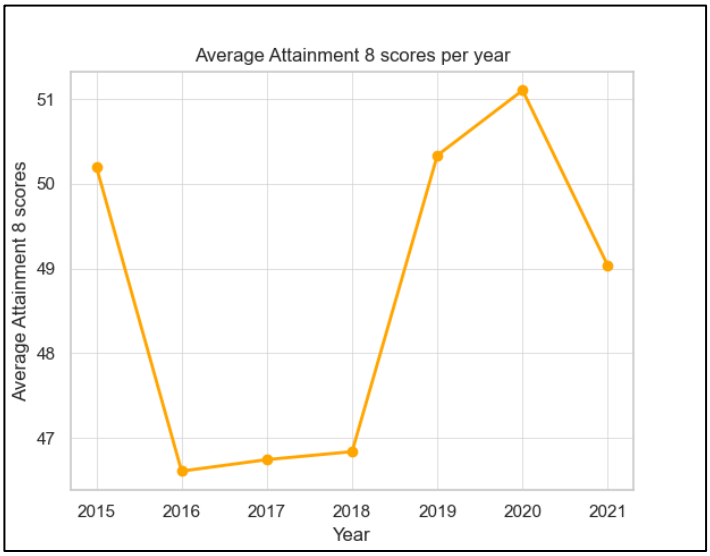


Figure 2:
Average Attainment 8 scores per year



Features

With Attainment 8 as our target variable, we compile a new dataset using Local Authority level data for England which might be relevant for predicting attainment. We used data that is published[1] on an annual basis by the Department for Education (DfE) as part of three publication themes, and we manipulate them using R.

Finance and Funding

The publications in this series provide us with information on the income and expenditure of schools. This details the amount of funding a particular LA received via different streams, and how schools spent that income on different categories of expenditure. We can understand the revenue balances of schools by combining this information on income and expenditure. We also have information on the amount a LA spent on other education and community services as well as children's services, which does not directly involve funding schools. For example, children's social care is funded in this way. We obtained 226 measures related to finance and funding.

Teachers and School Workforce

This publication[2] contains information on things such as the size of the school workforce in each LA, the types of teachers and support staff in the schools, and the qualifications of

[1] All the data used in this work is freely accessible at Explore Education Statistics (https://explore-education-statistics.service.gov.uk/data-tables)
[2] https://explore-education-statistics.service.gov.uk/find-statistics/school-workforce-in-england

teachers. It also includes information on teacher vacancies and absence rates, which are important for understanding the actual capacity of schools. The pupil-teacher ratio, a much-studied variable in the education literature, is also derived from this data. Finally, the publication also contains information on the pay of teachers, support staff, and other management staff in schools. We obtained 183 measures on the characteristics of teachers and the school workforce.

Pupils and Schools

The publications in this theme provide us with information on the characteristics of schools and the pupils that attend them. About the pupils themselves, we have data on their ethnicity and whether their first language is English, whether they are eligible for Free School Meals (which is usually an indicator of lower socio-economic background), their rates of absence, suspension, and exclusion, and any type of Special Educational Needs and Disabilities (SEND) they may have. These include the types of state-funded schools present in an LA, the admissions processes and class sizes in these schools, and the types of support in place for pupils with SEND. We obtained 134 measures on the characteristics of schools and the pupils that attend them.

Missing Values

In compiling this dataset, we inevitably encountered some missing data, which we dealt with in a few different ways. Circa 0.2% of the observations did not have Attainment 8 data, so we drop these from our sample as the small reduction in sample size is not a concern. For the funding and finance data, we replace any "NA"s with zeros. This is because, in the authors' previous experience working with the data, NAs usually represent Local Authorities that didn't receive any funding under a particular funding scheme or did not spend anything on items in a particular expenditure category. This reporting practice leads to the majority of observations being NAs for some features related to funding. For example, there were 14 features for which values were missing for over 60% of the observations. These features aside, the majority of features did not have a significant level of missing data (51% had less than 2% missing values, and 91% had less than 10% missing values).

For the teacher workforce and school and pupil characteristics data, missing values do not represent situations where it was truly "not applicable". We remove features that have a very high proportion of missing values. We choose to do this as they are unlikely to be informative even if we impute those missing values, and these 10 features only represent a small proportion (<2%) of the total number of features. On the remaining features, we impute the missing values using the median value for that particular Local Authority, not the median value across England. This means that for Local Authorities that have a missing value for the majority of the time series, such that the median value would be "NA", we don't actually impute the missing value (more precisely, the imputed value is still "NA"). We then drop these observations from the sample. We choose to do this because for example, if we observe the number of teachers in a Local Authority for all years except one, using the median value for that LA seems reasonable to impute the missing year, but for Local Authorities with a lot of missing data, this becomes less reliable. To preserve as much of the variation between LAs, using the national median value would not have been appropriate.

Throughout these procedures we have assumed that the data was missing completely at random. If this was not the case, then our chosen methods will impute values that are very dissimilar to the truth and introduce some bias in our data. Furthermore, the imputation methods we used only look at the distribution of the values of the variable with missing data. It's possible that the fact that Local Authorities fail to report on some data is correlated with some other characteristics in the data. However, we don't take advantage of this potential correlation to explore more complicated data imputation methods such as regression imputation or matrix completion (which uses principal components to impute missing values) as it is not the focus of our analysis.

School funding data is only reliably collected from 2015 onwards, with the latest available data being for the year 2021. Although we have school workforce, school and pupil characteristics, and attainment data for a longer time series, we limit our dataset to the years 2015 to 2021 for a more complete dataset. Once we implement the aforementioned procedures to deal with missing values, our resulting dataset contains 826 observations and 533 features in total. Since the number of features is close to the number of observations, we have high-dimensional data, which presents certain challenges that we will refer to throughout the analysis.

Methodology

We pre-process our dataset in Python by scaling all features such that they are centred around 0 and have a standard deviation of 1. This standardisation is necessary for certain methods we employ. We then randomly split the data and use 75% of it for training and the remaining 25% for testing purposes. We then begin the iterative process of finding the best performing predictive models. We start with four different tree-based methods. For each method, we use cross-validation to select the best performing model in that class. The performance of the model is assessed using the mean squared testing error, which is a measure of the difference between the model's predictions of the target and the actual target.[3] We also report on the testing R^2, which measures the fraction of variance explained, to offer an absolute measure of performance. By comparing across these two metrics, we identify the best performing tree-based method. We then apply Principal Component Analysis for dimensionality reduction purposes and see whether it improves the performance of tree-based methods by repeating the previous analysis. Lastly, we also employ linear methods, in particular penalised regression, to compare the performance of those models against tree-based methods.

Tree-based Methods

We approach our supervised regression problem starting directly with tree-based methods. These are non-linear methods, which means we do not have to assume a linear relationship between the features and the target unlike a method like Ordinary Least Squares. Given the large number of features, it would also become difficult to explicitly specify interaction terms in a linear model, whereas a decision tree naturally models the effect of one feature depending on the value of another as the tree is sequentially learned. A main drawback with applying decision trees to our problem is that we lose the main advantage of interpretability which is

[3] The formula is given by $MSE_{Te} = \frac{1}{n}\sum_{i \in Te}(y_i - \hat{f}(x_i))^2$, where n is the number of observations in the testing sample and $\hat{f}(x_i)$ is the predictions obtained by applying the estimated model to the testing data.

characteristic of decision trees. This is because of the high dimensionality of the data, which will likely allow trees to grow quite "deep" (with many nodes). However, a linear method also does not offer a simple interpretation when the number of features is very large. Therefore, we think interpretability of any model in our case is challenging so do not use it as a criterion for deciding between methods. We focus only on their performance in predicting Attainment 8 score.

We implement a decision tree, and use 5-fold cross-validation to select the optimal number of nodes. The best performing decision tree has 6 nodes and has a MSE of 0.41 and an R^2 of 0.64. This basic level of performance is what we use to benchmark against when we explore more advanced ensemble learning methods. Decision trees are prone to overfitting, or in other words suffer from high variance. The idea behind ensemble learning is to take weak learners, which are smaller models such as our decision tree which taken singularly produce mediocre results, and aggregate them into a single large model that may produce better forecasts. Bootstrap aggregation (or bagging) works by taking *B* repeated samples of the training data, constructing *B* trees, and then averaging the forecasts. Random Forest tweaks this process by only considering a subset of all the predictors for each split when building the decision trees. This process of decorrelating the trees works by avoiding one very strong predictor in appearing as the top split in most bagged trees, which means they'd end up looking very similar and be more correlated. It can be argued that Random Forest offers an improvement over bagging when it's observed that the predictors are highly correlated, as the trees will likely also be more correlated in that situation. Given the dimensionality of the dataset, it is difficult to visually inspect a correlation matrix or a heatmap. We check the degree of correlation between predictors by calculating the proportion of predictor pairs that have a correlation coefficient above 0.7. This turns out to be 3.71% (5,119 out of a total 138,075 pairs), which leads us to conclude the data is not very highly correlated, so a Random Forest may not have a big advantage over bagging.

Considering that bagging is just a special case of Random Forest where all the features are available for each split when constructing the trees, we implement both methods in a similar way. Using 5-fold cross-validation to tune the hyperparameters, we have a bagging regressor which averages the forecasts of 75 trees with 15 nodes each. This is a complex model that makes interpretability difficult, but the model achieves a good level of performance with a MSE of 0.22 and an R^2 of 0.80, which is a significant improvement over the simple decision tree. A Random Forest also provides similar levels of performance, with a MSE of 0.25 and an R^2 of 0.77. The cross-validated random forest estimator has 50 trees, each with 15 nodes. Our earlier hypothesis that, since the features are not highly correlated, a Random Forest may not have a big advantage over bagging was borne out, in fact Random Forest performs slightly worse. Being able to consider all features at every split led to an improvement in performance, which could be an indication that there aren't many particularly strong predictors of the target but rather many somewhat relevant features.

Both of the ensemble methods above fit a single large tree to the data, which potentially amounts to overfitting, so we explored boosting as an alternative. Boosting involves growing trees sequentially by fitting a (usually small) decision tree to the residuals of the previous tree with the idea to improve the fitted function in areas where it doesn't perform well. With a slow learning rate, boosting tends not to overfit. We therefore fit a gradient boosting regressor and

use 5-fold cross-validation to tune the hyperparameters. A gradient boosting regressor with a learning rate of 0.01 and 200 trees of 4 nodes each results in a MSE of 0.26 and an R^2 of 0.77. This is similar to the performance of the random forest, but nonetheless is slightly worse. This leads us to think overfitting may not be the principal limitation to the performance of our random forest model.

Table 1: Performance of tree-based methods

Model	**Mean Squared Error (MSE)**	**R^2**
Decision Tree	0.41	0.64
Bootstrap Aggregation	0.22	0.80
Random Forest	0.25	0.77
Gradient Boosting	0.26	0.77

Principal Component Analysis

Principal Component Analysis is an unsupervised technique, which means it is not used for prediction in itself. In our case, we want to use it for dimensionality reduction purposes. The idea is to create a small number of linear combinations of the features (the principal components) that explain most of the variance of the data, then train a model on those principal components (PCs). Since the dimensionality of that data should be lower (number of observations remains unchanged and a small number of PCs are used instead of 533 features), this may reduce the overfitting problem and improve prediction performance.

Figure 3: Cumulative Variance Explained by Principal Components

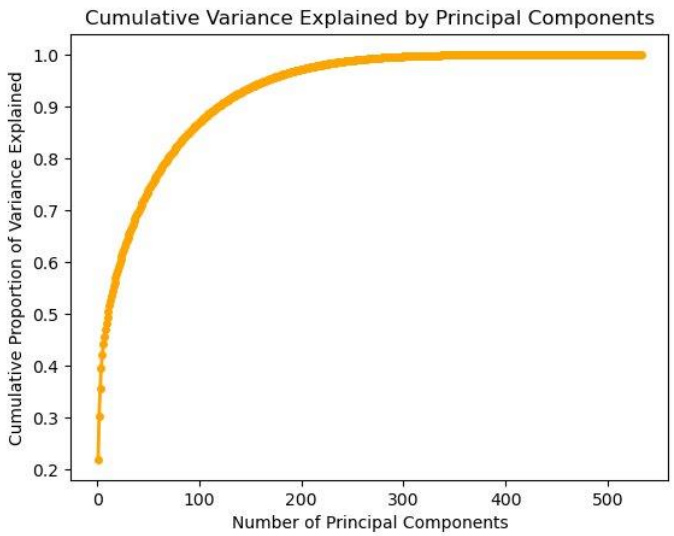


Figure 3 plots the number of PCs against the cumulative variance explained by them. This shows we do not have a case where a few PCs explain most of the variance. The first 10 PCs explain roughly 50% of the variance in the data. There is also no obvious "elbow" in the plot, which indicates the additional proportion of the variance explained by the subsequent PC is much smaller. The smooth curve doesn't provide us with an easy decision point on how many PCs to select. Instead, we determined the number of PCs that would explain 90% of the variance, which we deem to be "most" of the variance, to be 119 PCs and use those in our analysis.

We repeat our previous analysis, and fit a decision tree, a random forest estimator, and a gradient boosting regressor to the 119 PCs. Again, we use 5-fold cross-validation to determine the best performing model of each method. Table 2 below reports on the performance of these models. The main conclusion is that each model performs worse than their counterpart trained on the original data.

Table 2: Performance of tree-based methods on the Principal Components

Model	**Mean Squared Error (MSE)**	**R^2**
Decision Tree	0.80	0.29
Bootstrap Aggregation	-	-
Random Forest	0.39	0.65
Gradient Boosting	0.49	0.56

Against conventional wisdom, PCA hasn't improved performance in this case. There are a few possible reasons for this. The number of components used is still very large (119), and whilst it is certainly less than the original number of features (533), it does not reduce the dimensionality of the data to an extent that we would consider it low dimensional data. 10% of the variance, which could contain information useful for prediction, is lost in this process as well. The trade-off posed by PCA might not be worthwhile in this case, as the tendency for decision trees to overfit is not avoided but information is lost regardless. This situation is likely due to the signal to noise ratio in our dataset being quite low. Meaningful PCs can be produced only when the signal to noise ratio in the data is high, meaning that many features are truly associated with the target (signal) rather than having chance associations (noise). Using these meaningful PCs would then be effective at dimensionality reduction and reduce the testing error of our models by reducing the risk of overfitting.

Penalised Regression Methods

Our decision to apply nonparametric methods such as decision trees from the start was driven by their greater flexibility which we thought would lead to improved prediction performance. Nonetheless, for completeness we also explore some linear methods, particularly penalised regression methods. To be clear, these methods are used on the original data, not the principal components.

We start with a regression using the Least Absolute Shrinkage and Selection Operator (Lasso). This method adds a penalty term to the Ordinary Least Squares (OLS) objective function, where the penalty is the absolute sum of the coefficients multiplier by a parameter (alpha), which encourages sparsity by shrinking some coefficients to zero. We use 5-fold cross-validation to choose alpha, which ends up being 0.1. This is surprising, as larger values of alpha tend to achieve the lowest test error in a high-dimensional setting. With the lasso, we have a MSE of 0.38 and an R^2 of 0.66, which is comparable to the performance of a basic decision tree but substantially worse than the ensemble methods. In the lasso, 515 features have coefficients equal to exactly zero, which means the model only used 18 features to form predictions. In performing this feature selection, the oversimplification may have introduced too much bias[4].

The ridge regression also includes a penalty term in the objective function, but this time it is the squared sum of the coefficients multiplied by alpha. This has the effect that the coefficients are shrunk towards zero but do not reach zero. This means it is likely to produce estimates which are more stable by reducing the variance at the expense of not performing any feature selection.

[4] The fact that the training error is similar in magnitude to the testing error supports this idea.

Using 5-fold cross-validation, we set alpha to be 100, which indicates a high level of regularisation[5]. This regression has a MSE of 0.17 and an R^2 of 0.85, which is surprisingly good. In this case, the reduction in bias incurred in having many coefficients not be exactly zero offsets the increased variance.

We also apply the elastic net penalty term, which essentially combines the lasso and ridge penalty terms, to see if it can offer even further improvements. The cross-validated elastic net model has a MSE of 0.19 and an R^2 of 0.83. Using 5-fold cross-validation, the best value for both of the two hyperparameters, alpha and the "l1_ratio", is 0.1. The value of alpha is quite small, which results in weaker regularisation and allows the model to fit the training data more closely by not pushing the coefficients of the features more towards zero. When applying the elastic net, if the l1_ratio is 0 only the ridge penalty is used and if the l1_ratio is 1 only the lasso penalty is used. An l1_ratio of 0.1 indicates that the elastic net more closely resembles the ridge regression than the lasso regression. However, it performs slightly worse than the ridge because loosely speaking, it still has some element of the lasso regression which overall performs worse. The combination of a low alpha and a l1_ratio close to 0 indicates that the elastic net is not doing much feature selection (which is what the lasso would do).

Table 3: Performance of penalised regression methods

Model	**Mean Squared Error (MSE)**	**R^2**
Lasso	0.38	0.66
Ridge	0.17	0.85
Elastic Net	0.19	0.83

Both the elastic net and ridge regression outperform ensemble methods, with the ridge regression performing best overall. This is again a surprising result, and we offer a few hypotheses as to why this is the case. Firstly, it could be the case that the relationship between the features and the target is primarily linear, so there is less value in relaxing the linearity assumption. The earlier discussion on signal to noise ratio continues to apply in highlighting the benefits of regularisation given the high dimensionality of the data. It could be that a ridge regression is better suited at dealing with the risk of overfitting than the bagging model given it is less flexible. By comparing the R^2 obtained on the training set for both methods, we see that the bagging model achieves a training R^2 of 0.96 and the ridge regression achieves a training R^2 of 0.91. Even though the bagging model fits the training data better, it generalises less well than the ridge regression, suggesting the bagging model suffers more from the issue of overfitting. Linked to this issue of high dimensionality is the consideration that tree-based models tend to do better when there is a large amount of training data available to capture complex relationships, and maybe our data was not sufficiently large to make full use of these methods.

Model Inspection and Policy Recommendations

Having concluded that the ridge regression is our best performing model in terms of predicting Attainment 8, we want to understand which features are important in this model. As a

[5] Regularisation is any modification we make to a learning algorithm that is intended to reduce its generalisation error but not its training error (Goodfellow et al., 2017).

standardised linear model, we just need to look at the size of the absolute value of the model coefficients to identify the features that have the largest impact on Attainment 8. This interpretability is an important advantage of the ridge model over the ensemble methods. Using techniques like permutation importance to inspect ensemble models would only allow us to understand the most important features for a particular model, rather than the intrinsic predictive value of those features[6]. We report on the coefficients of the ten most important variables for the ridge regression in Annex A. We can sort the ten most important features into four distinct groups that we believe policymakers should focus on.

Inequality

We find that two of the top ten most important features are related to the ethnic background of pupils. Specifically, these are the proportion of pupils in a Local Authority from Bangladeshi backgrounds and Mixed backgrounds. Both of these features have a positive effect on Attainment 8. We also find two variables related to Special Educational Needs and Disabilities (SEND) to be important and both have a negative effect on Attainment 8. These features are the number of new Education, Health, and Care Plans (EHCPs) and whether a Designated Clinical Officer is in place in the Local Authority. Both of these are forms of additional support for pupils with SEND so are indicators of the prevalence of SEND in an LA, and their negative effect on Attainment 8 can be interpreted as pupils with SEND performing worse than their peers despite the additional measures. These point to the existence of systematic inequalities existing in the education system, a finding echoed by the IFS Deaton Review (2022). Although reducing these inequalities is not entirely within education policymakers' control, it should not discourage implementing initiatives to that effect.

Absence and Suspensions from School

Another feature in the top 10 is the suspension rate of pupils, which unsurprisingly has a negative effect on Attainment 8. A further three features are related to the authorised absence rate of pupils, which have a negative effect on Attainment 8, unless the absence was authorised for religious reasons, in which case it has a positive effect. A report by the DfE (2015) suggests that those who had the lowest 5% of absence rates were 4.6 times more likely to achieve more than five A*-C grades including English and Maths compared to those who had the highest 5% of absence rates. Evidence shows that improving communication with parents via low-cost technology is an inexpensive and effective intervention for addressing this issue (Sanders et al., 2019). It is evident that policymakers can do more to improve absence rates, but further investigation into the underlying mechanisms of different measures of absence is needed, as often absences are reflective of larger underlying challenges like mental health or family issues.

Targeted Grants

The only variable related to funding to feature in the top 10 is the amount of targeted grants, which has a positive effect on Attainment 8. This could be interpreted as showing funding which is targeted (i.e. delivered to address a specific issue in the Local Authority) can be an effective way to address shortfalls or distortions associated with the general funding system and improve attainment results. Nicollete and Rabe (2012) look at the relationship between grants per

[6] See Annex C for a chart of the most important features in the bagging model.

student and test scores that they achieve when they are 16. The paper concludes that a £1,000 increase in spending per student will results in a 0.2 standard deviation increase in Attainment 8 scores. However, we find a limited number of academic studies that present any form of conclusion for policymakers on the best way to fund schools.

Teachers

The only variable related to teachers to feature in the top 10 is the number of teachers with at least degree-level qualifications (including a Bachelor of Education and a Postgraduate Certificate in Education). This can be interpreted as an indicator of the quality of the teachers, and unsurprisingly it has a positive effect on Attainment 8. Unfortunately, England has a severe teacher recruitment shortage across most subjects (McLean et al., 2023). Retention bonus payments have been shown to have a large impact on retention (Sims and Benhenda, 2022) and so could be an effective means of recruiting more teachers. However, policymakers might consider creating additional incentives specifically for teachers with at least degree-level qualifications by targeting retention bonuses.

Conclusion:

We have used a high-dimensional dataset to build a predictive model of Attainment 8 scores. We included data at Local Authority level on school finances, the school workforce, and the characteristics of the pupils in our dataset. We explored the use of tree-based methods, and observed that various ensemble methods offered a reasonable prediction accuracy. We then turned to an unsupervised machine learning method in the form of Principal Component Analysis to summarise the data and reduce its dimensionality. The aim of this was to reduce the risk of overfitting in our models. However, because of the low signal to noise ratio in our data, the dimensionality reduction was not effective, and we did not improve the prediction performance of any of the tree-based methods. Although tree-based methods, particularly ensemble methods, are often considered to be the best choice of model, we still explored the use of linear methods. In particular we applied three different penalised regression methods, including the Lasso regression, ridge regression, and the Elastic Net. To our surprise, the ridge and elastic net regressions outperformed even the ensemble learning methods, with the ridge regression having the highest prediction accuracy with a testing R^2 of 0.85.

In building the predictive model, we have also helped uncover some important factors which are driving differences in educational attainment across Local Authorities. These can be grouped into four distinct categories, which are inequalities, absences from school, targeted funding, and teacher quality. We have made a few policy recommendations for how to address these issues, whilst acknowledging that due to the existence of financial constraints there is no "low-hanging fruit" to improving attainment. All the factors we identified to be important for predicting attainment are areas where the education sector is currently facing system-level challenges, leaving plenty of challenges for policymakers to grapple with.

References

Allison, Paul D. Missing data. Vol. 136. Sage publications, 2001

Baker, D., Goesling, B., & Letendre, G. (2002). Socioeconomic status, school quality, and National Economic Development: A cross-national analysis of the "Heyneman-Loxley effect" on mathematics and Science Achievement. Comparative Education Review, 46(3), 291. https://doi.org/10.2307/3542092

Bengio, Y., Goodfellow, I., & Courville, A. (2017). Deep learning (Vol. 1). Cambridge, MA, USA: MIT press.

Cheti Nicoletti & Birgitta Rabe, 2012. "The effect of school resources on test scores in England," Discussion Papers 12/19, Department of Economics, University of York.

Department for Education (2015), 'The link between absence and attainment at KS2 and KS4'. https://assets.publishing.service.gov.uk/media/5a802a2d40f0b62302691e66/The_link_between_absence_and_attainment_at_KS2_and_KS4.pdf

Farquharson, C., McNally, S. and Tahir, I. (2022), 'Education inequalities', IFS Deaton Review of Inequalities.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani and Jonathan Taylor. (2023). An Introduction to Statistical Learning: with Applications in Python. New York: Springer.

McLean, D., Worth, J. and Faulkner-Ellis, H. (2023). Teacher Labour Market in England: Annual Report 2022. Slough: National Foundation for Educational Research.

OECD (2019), PISA 2018 Results (Volume II): Where All Students Can Succeed, PISA, OECD Publishing, Paris, https://doi.org/10.1787/b5fd1b8f-en.

Sanders, Michael and Kirkman, Elspeth and Chande, Raj and Luca, Michael and Linos, Elizabeth and Soon, Xian-Zhi, Using Text Reminders to Increase Attendance and Attainment: Evidence from a Field Experiment (March 8, 2019). Available at SSRN: https://ssrn.com/abstract=3349116 or http://dx.doi.org/10.2139/ssrn.3349116.

Sims, S., & Benhenda, A. (2022). The effect of financial incentives on the retention of shortage-subject teachers: evidence from England (CEPEO Working Paper No. 22-04). Centre for Education Policy and Equalising Opportunities, UCL.

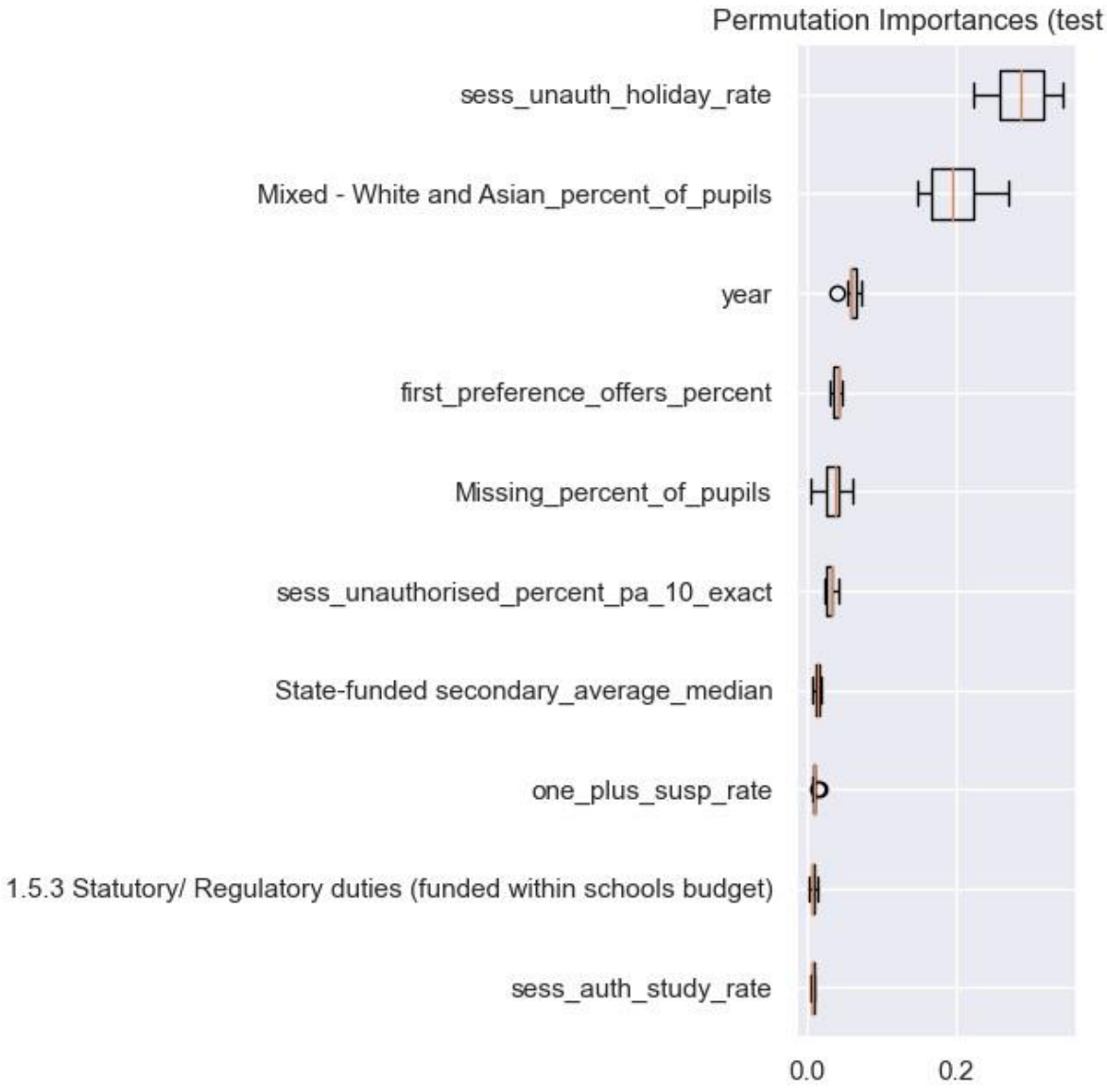
Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67: 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Annex A: Results of the Ridge Regression Model

Table A: 10 most important features

Feature name	Coefficient
Asian - Bangladeshi_percent_of_pupils	0.123
Total targeted grants	0.109
Sess_auth_totalreasons_rate	-0.098
Sess_authorised_percent_pa_10_exact	-0.089
Sess_auth_religious_rate	0.084
Num_new_EHCP	-0.079
Designated_clinical_officer_in_place	-0.077
Susp_rate	-0.076
Degree or higher / Bachelor of Education / PGCE_headcount	0.073
Mixed - Any other Mixed background_percent_of_pupils	0.073

Annex B: Inspecting the Bagging Model using Permutation Importance



School of Economics and Finance

Queen Mary University of London

This working paper is based on project work undertaken by EMAP apprentices

School of Economics and Finance
Queen Mary University of London
Mile End Road
London E1 4NS
Tel: +44 (0)20 7882 7356
Fax: +44 (0)20 8983 3580
Web: www.econ.qmul.ac.uk/research/workingpapers/