# How can DfE most effectively contribute to improving outcomes for disadvantaged pupils in England?

Dissertation Report

## Contents

Word count (abstract & main body): 14,447

- ## Abstract

The Department for Education holds a commitment to improving outcomes for disadvantaged pupils as one of its core objectives. The ways in which the department can most effectively achieve this goal are not well understood. Most prior studies have been hampered by difficulties in accounting for complex interaction effects, a limited ability to link government policy to labour market outcomes, and insufficient statistical power to detect small effects. This study contributes to addressing these issues by implementing a novel methodology combining machine learning and traditional statistics, and by utilising new large administrative data sources linking pupils' education data with earnings later in life. The analysis is conducted in two stages. In the first stage, we build a gradient-boosted, tree-based machine learning model, tuned with a Bayesian hyperparameter optimisation procedure, to predict what pupils' early-career salaries will be, six to eight years in advance. We use this model to explore interaction effects in the data, and to generate data-driven hypotheses about the types of policies schools implement that seem to be associated with better labour market outcomes for disadvantaged pupils. The model has a moderate level of predictive power ($R^2$ = 0.37; RMSE = 7630). In the second stage, we test these hypotheses on new data using a Penalised Quasi-Likelihood mixed-effects model, with a school-level random intercept. We find that pupils' performance in Maths at KS2 has the strongest effect on early-career earnings, out of the independent variables tested ($\beta = 1137$, std. error = 81, $p$ < 0.0001). Crucially, this effect is moderated by gender and economic disadvantage: the effect of maths performance on earnings is significantly larger for girls than boys and for disadvantaged vs affluent pupils. The strength of the interaction is such that, for high-performing pupils, the effect of disadvantage is entirely eliminated. Finally, we find evidence for a moderate positive effect of KS4 cohort size (average-sized cohorts are best), ethnic diversity, and recent capital investment in school improvements. Policy implications and suggestions for future research are discussed.

# • Introduction and rationale

Improving social mobility is one of the core stated objectives of the Department for Education (DfE, 2017). The consensus in the education research literature is that schools are likely to provide the best vehicle through which the department may be able to achieve this aim (see Burgess, 2016, for a review). Despite this, the link between schools' activities – the approaches that heads, school leaders and teachers take to improve outcomes for their pupils – and long-term social mobility outcomes is not fully understood.

Given that schools account for the majority of the department's £60bn+ budget (DfE, 2016), and the Government's belief that education is a central component in improving social mobility, it is clear that the question of how to use schools to maximum effect in achieving this goal should be of great interest to policymakers.

## Originality, objectives, practical application

Despite this interest, there is a gap for a project using innovative methodologies to explore this question directly. There are three main reasons for this:

1. Most existing research focuses solely on school-age academic performance as the outcome of interest (see Burgess, 2016)
2. The majority of this research does not focus on disadvantaged pupils, or on complex interactions between disadvantage and other factors
3. Where quantitative methodologies are employed, almost all have used traditional statistical methods, and many have suffered from an inability to unpick complex interaction effects. Only one paper we are aware of has used machine learning techniques (for example), and this included few factors schools could control, and did not focus on disadvantage (Masci et al, 2018)

This project will therefore bring an original perspective to the topic by (a) investigating the effect of schools' activities on labour market outcomes directly, (b) focusing on disadvantaged pupils, and (c) utilising innovative approaches that have not yet been widely employed in the literature.

The aim of this project is to:

a) develop a **predictive model** that quantifies the effect of school behaviours on disadvantaged pupils' labour market performance after leaving school, controlling for key pupil and school background characteristics, and

b) to develop and **test hypotheses** about how schools can most effectively influence their disadvantaged pupils' long term outcomes.

Practically, we will then use this model to contribute to the department's work on improving outcomes for disadvantaged children. Possible applications could include:

- Influencing DfE policy – e.g. school improvement funding
- Contributing to DfE guidance to schools, e.g. on the use of the Pupil Premium
- Developing a tool for headteachers to use to directly access the findings and receive bespoke, data-driven advice on how best to improve results for disadvantaged pupils in their school

## Background review

### Definition of terms

Given our overarching question ("how can schools most effectively contribute to social mobility in England"), we have two key areas to define: schools (and what they can "do"), and social mobility.

### Schools

As DfE is not responsible for education policy in Wales, Scotland, or Northern Ireland, we will focus our attention purely on schools in England.

There are many different dimensions along which to divide the population of schools. Two of the most important are:

- Academies vs maintained
    - Maintained schools are what most people would think of as traditional state schools. They are funded and controlled by the local council.

    - Academies are “publicly funded independent schools”[1], run by trusts or sponsors. They are funded directly by central government, rather than the local council, and have more autonomy than maintained schools in making various decisions (such as spending and curriculum choices).
    - We include both types of school in our analysis

- Primary vs secondary
    - Primary schooling lasts from ages 5 to 11, and secondary lasts from 11 to 16
    - There is extensive research on the relative importance of the two phases of education for pupil attainment (*Psacharopoulos & Patrinos, 2018)*. Most studies tend to conclude that earlier intervention is more influential, with greater economic returns for primary than for secondary schooling
    - In addition, we know that the performance gap between disadvantaged pupils and their more affluent peers can already be observed at KS2 (EEF, 2018)
    - We will include information about both primary and secondary schooling  in our analysis, to account for the effects of different school activities at different points in a child’s life course.

The independent variables included in our analysis fall into two groups::

- **Control variables** – those that schools largely cannot influence. Examples include the characteristics of the school, the pupil, the pupil’s peers, and the school’s local area.

- **Independent variables** – those that schools *can* have some influence on. These relate to the “activities” that schools undertake: their approaches to teaching, the spending choices they make, and so on. These include spending, leadership, teacher recruitment & retention, and so on.

**Social mobility**

There are many possible definitions of social mobility used in different contexts (Crawford et al, 2011). Given that we are focused on practical application for DfE policy,

[1] https://www.gov.uk/types-of-school/academies

we will employ the DfE definition for our research. In this view, social mobility reflects the ability of children from poor backgrounds to achieve well-paid jobs later in life (DfE, 2017). A perfectly socially-mobile society is therefore one in which someone from a poor background is equally likely to secure a well paid job later in life as somebody from a wealthy background.

There are, of course, limitations to this definition, including:

- Income is a crude measure of one's position in society. Social, cultural and human capital (e.g. Farkas, 1996; Baron et al., 2000) all play a significant part in affecting an individuals' life experiences (although researchers disagree on the relative importance of each). This limitation is mitigated to some extent by the fact that the DfE definition implicitly includes a focus on human capital (which is generally defined in terms of education and skills), but cultural and social capital are rarely considered as key aims for DfE policy.
- Above a certain level, income is a poor measure of life satisfaction, with most studies finding significant diminishing returns to increased income (e.g. Diener et al, 1993; Diener & Biswas-Diener, 2002)
- Societies in which all individuals have a very low, but equal, chance of achieving a well-paid job would score well on this metric, even though this is clearly an undesirable outcome. It would therefore be a poor measure in certain circumstances, such as for developing countries

However, complex terms like "social mobility" naturally resist classical definition, and all such definitions will be imperfect. Influencing the Government's definition of social mobility is beyond the scope of this project; we will therefore adopt this definition in order to maximise the relevance of our research for our users (DfE policymakers).

### Income

Our dependent variable relates to pupils' incomes (or annual earnings) 'later in life'. The earnings data we have available comes from the Longitudinal Earner Outcomes (LEO) database, which matches pupil records to HMRC data on earnings. The latest year available is from 2017, and includes individuals up to the age of 30. This means that the oldest individuals for whom we have earnings data completed their KS4 exams around

2002. However, the oldest year for which we have complete data for our chosen independent variables is 2009.

Thus our selected cohort for this project comprises pupils who have **both** complete data for our independent variables **and** records in the LEO database. This restricts us to the cohorts that completed their KS4 exams in 2009-2011, whose latest LEO record is from 2017, when they were aged around 22-24. Our dependent variable is therefore: **earnings 6-8 years post KS4**, excluding those not in employment.

### Research question

We can use these definitions to operationalise the key terms in our overarching question and create a more clearly defined (albeit slightly more unwieldy) research question:

What are the characteristics and behaviours of state-funded schools in England that have the strongest influence on the salary of pupils from poor backgrounds 6-8 years after leaving school?

## • Literature review

The predominant logic model in the literature on effects of schooling on disadvantaged pupils' labour market outcomes can be formulated as:

- Schooling → skills & qualifications → labour market outcomes

We will review the literature on the two halves of this relationship separately.

### Schooling → skills & qualifications

The first major finding in the literature on this area is that school-level effects on skills and qualifications are **small** and **complex**, with the majority of variance explained at the pupil level (e.g. Coleman et al., 1966; Hanushek, 2008). There is more variation in pupil-level performance *within* schools than *between* schools, and it appears to matter more who you are than where you go to school. This is perhaps surprising, and is probably at odds with the public perception that going to a "good school" can drastically alter your life chances.

However, it may be that these effects only appear small because of the complex interactions involved in the production of human capital. Very few researchers have attempted to unpick these using methods that specialise in complex interactions. It is perhaps telling that the only study we know of to have done so found larger than average effects for school-level factors (Masci et al 2018)

The second major finding from this body of research is that the relationship is moderated by disadvantage. Disadvantaged pupils make less progress on average during their school careers than non-disadvantaged pupils – although there is significant overlap (e.g. Allen, 2018). At the same time, disadvantaged pupils also appear to benefit more from school-level factors (e.g. particularly good teaching) than non-disadvantaged pupils (EEF, 2017).

How do we resolve this apparent contradiction?

It seems that being economically well-off is a kind of insulating factor against the influence of school. For children from affluent families, which school the child attends has a *relatively* small impact on educational outcomes – possibly because the child will benefit from interventions like private tuition, extra resources, extra help and motivation from home life regardless of the effectiveness if the school. Indeed, most studies investigating the home learning environment (HLE) have found it to be a stronger predictor of academic performance than schooling (Lessof et al., 2018). Thus when a child from an affluent family attends a less effective school, their family's resources fill the gaps in their education.

Conversely, disadvantaged pupils are more sensitive to school influences. They are less likely to get their motivation to learn from home, less likely to have dedicated space and resources for learning at home, and may even be less likely to have basic needs met outside of school (Lessof et al., 2018). They are therefore more in need of school to provide all of this. This means that if the school *does* provide it, it makes a big difference (unlike for better-off pupils), and if the school doesn't then performance drops off considerably. In turn, this suggests that disadvantaged students simply aren't getting the support they need to fulfil their potential (support that well-off pupils get outside of school).

Taken together, these findings suggest (a) that our research may find larger effects for school-level factors if we choose methods appropriate for unpicking complex interactions, and (b) that the impact of these effects could be substantial for the most disadvantaged pupils (i.e. those we most want to help), despite the *overall* small effect sizes.

## Skills & qualifications → better-paid employment

There is a wealth of literature on this half of the relationship (e.g. Bhutoria, 2016, and references). In the UK, there is a very strong link between formal qualifications and income. There is some debate over the extent to which this results from actual accumulation of skills vs "screening" effects (where employers are simply biased in favour of those with more qualifications, even if they are irrelevant for the job at hand). However, studies comparing people who narrowly passed their qualifications and those who narrowly failed have shown that the two groups tend to have similar employment outcomes. This suggests that the screening effect is at most minimal, and that academic qualifications – although imperfect – do a reasonable job at measuring aptitude for work at the point of entry into the labour market.

As with the effect of schooling on the accumulation of human capital, this relationship is moderated by disadvantage. Disadvantaged pupils appear to benefit more from qualifications than non-disadvantaged pupils (Sharp et al, 2015), but their more affluent peers tend to be more likely to get highly-paid jobs even if they have the same academic qualifications (Crawford et al, 2016). This may result from factors like unconscious class bias, cultural capital, and social networks.

Are there any activities that schools can engage in which would help disadvantaged pupils to do as well as their more affluent peers in the labour market, over and above teaching and qualifications? Schools certainly undertake a number of such activities (e.g. careers advice, mentoring, mental health & wellbeing coordinators and counsellors), but there is surprisingly little research on the effectiveness of these activities.

## Issues & considerations

There will be some legal considerations to be taken when conducting the project. It will involve the use of some potentially sensitive personal data relating to children, taken from the DfE's Pupil Data Repository. This will need to be linked to other school-level data (such as the financial returns on school spending). The project will comply with all relevant legal restrictions, including those covered by GDPR, by utilising only data that is necessary for the analysis, conducting all analyses on secure DfE networks, and using the results only for uses that schools, pupils and parents have consented to (such as for anonymised research purposes and for guiding policymaking).

Ethical issues will also be considered. For example, we will take care when building the model that we are not inadvertently encoding societal prejudices and stereotypes (as has previously occurred in, for example, the US justice context - Israni, 2017). Such issues will also be considered when using the research. In particular, we would need to be careful around publication of the results, and preventing any predictions (e.g. around pupil characteristics) from becoming self-fulfilling prophecies, reinforcing stereotypes and causing harm in society.

### Quantifiable objectives

The quantifiable outcomes that we will produce for this project are as follows:

1. A set of exploratory visualisations on pupil earnings, pupil background, and school characteristics
2. A **predictive** machine learning model to predict pupil earnings 6-8 years post-KS4
3. A set of data-driven hypotheses about the **causes** of early-career income for disadvantaged pupils, generated using the results of the predictive model
4. A statistical model that **tests** these hypotheses on new data
5. An answer to the headline research question

## • Methodology

This section lays out the plan we formulated for our analyses, the justification behind the analytical methodologies employed, and a description of the development process, evaluation and results of each stage of analysis.

### Project planning

We divided our projects into two parts: data collection and analysis. The analysis stage was further divided into two stages - machine learning and statistical modelling. The outline project plan is shown below.

1. **Data collection**

    a. Extracting data from the schools census, KS2, KS4, Ofsted and LEO databases with SQL
    b. Matching the datasets with SQL
    c. Cleaning and filtering the data in R

2. **Analysis part 1**
    a. Exploratory data analysis
    b. Dimensionality reduction
    c. Model training on the reduced dimension set
    d. Model testing & evaluation
    e. Model training on the full feature set, for comparison
    f. Model testing & evaluation
    g. Variable importance extraction
3. **Analysis part 2**
    a. Hypothesis generation
    b. Model specification
    c. Run model
    d. Evaluate & interpret model

## Methodology selection & justification

The justification for this this two-part analysis plan runs as follows:

- Ultimately we are interested in both causal inference *and* prediction. We want to know what *causes* some disadvantaged pupils to perform well in the labour market (so that, as a department, we can do more of it), and we also want to be able to *predict* which cohorts are likely to have the poorest outcomes, so we know where best to target out efforts
- We therefore decided to divide the analysis into two parts: prediction and causal inference
- For the prediction task, our question naturally falls into the *supervised regression* category. That is, we have a set of quantitative predictor variables and a quantitative, continuous outcome variable (earnings), and we want to find a function that maps the former to the latter with as little error (or loss) as possible. Further, we want to find a function to perform this task as effectively as possible on **unseen data**.

- *Machine learning* modelling represents the current state of the art in the field of out-of-sample predictive modelling for continuous variables, so this is the broad category of models from which we selected our methodology.
- There are many, many types of model that can be used for this kind of task (see Singh, Thakur and Sharma, 2016, for a review). These range from simple probability models (e.g. naive bayes), basic nearest-neighbour classifiers (e.g. k-nn), and easily-interpretable tree-based methods (e.g. decision trees), to more complex and powerful methods for handling noisy data (e.g. randomforest, boosting, bagging etc), and the somewhat black-box neural network family.
- We decided to implement a **gradient boosted regression tree** methodology for this stage of our analysis, using the XGBoost R package (see Friedman, 2001 for an overview of the technique). We chose this methodology for several reasons:
    - Our data was relatively high-dimensional for a social-science context (250+ predictors), and we anticipated that there would be a significant amount of noise in the data as a result. The implementation we chose includes various methods to avoid **overfitting** in these kinds of situations, and thus maximise out-of-sample performance. These include: creating an ensemble of many shallow, weak trees to broaden learning; randomly subsetting the columns and rows available to train each learner, to avoid fitting to particular outlier cases or noisy features; and L1 &L2 regularization to reduce the weight of noisy features in formulating predictions
    - This kind of model fits trees iteratively and learns which features predict misclassification by the previous tree. Gradient descent is used to intelligently select the specification for each subsequent tree. This process is one approach to maximising predictive power, and although performance depends on the specific data under consideration, this family of models tends to perform very well on complex non-linear prediction tasks (e.g. in the leaderboards on Kaggle, a machine learning competition site). We anticipate that there may be many complex interaction effects in our dataset. For example, special educational needs (SEN) has a strong effect on

pupil attainment, and the quality of a school’s provision for SEN pupils is likely to have a large impact on attainment if the school has many SEN pupils, and a smaller effect if there are very few SEN pupils. There are many such possible interaction effects uncovered in the literature, and we wanted to take these into account in a data-driven manner, without specifying them all in advance.

- As well as prediction, we wanted to use the machine learning phase as an aid to generating a causal hypothesis. The XGBoost implementation includes an option for extracting feature importance, which gives each feature a set of scores relating to how useful they are for the model in creating its predictions. Our plan was to use these scores to help us generate our causal hypothesis.
- This is the approach implemented in Masic et al (2018), and so we are following an approach that has been successfully trialled in the education research literature

- For the *causal inference* task, a machine learning model was unlikely to provide what we needed. Machine learning models are generally designed for *prediction* rather than inference. As such, they are optimised to identify combinations of features that tend to coincide with specific values of the outcome variable, regardless of whether they are causally related. The second task therefore fit more naturally in the realm of inferential statistics.
    - Again, there were many methodological approaches available within this category.
    - We opted to implement a **mixed effects multi-level model** to test our hypothesis.
    - We did so for two reasons. Firstly, the structure of our data is inherently hierarchical, as pupils are naturally clustered within schools, thus violating the independence assumption for simple linear modelling. Secondly, the literature suggests that the majority of the variation in academic attainment and income is explained at the pupil level - i.e. factors relating to individuals’ motivations and backgrounds are more strongly predictive of outcomes than factors relating to their school are. Thus we decided to

include a random effects term for pupil identifier, so as to control for unobserved pupil characteristics when testing for the effect of the school features included in our hypothesis.

The analytical development conducted in parts 1 and 2 are discussed below.

## Part 1: Machine Learning

The first analysis phase was further broken down into four broad sub-phases: (1) exploratory data analysis, (2) dimensionality reduction, (3) training & testing a model on the reduced feature set, and (4) training & testing a model on the full feature set.

### Problem definition

We formulated our task as a *supervised regression* problem, of the form $y = f(X)$,

where y is the continuous earnings variable, *X* is the matrix of predictor variables, and f() is the function we are attempting to model. Before running our modelling, we first conducted some exploratory visualisation analysis on the data.

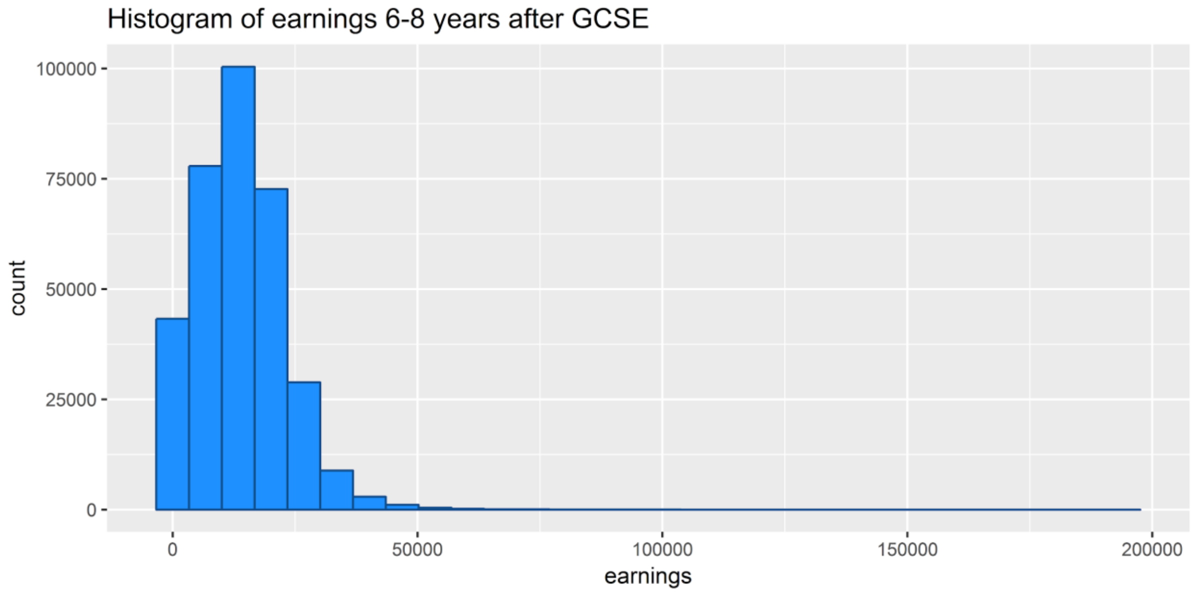
### Exploratory data analysis

All of the analyses were implemented using R, with code written and run in the R Studio IDE (R Core Team, 2017). Data manipulation and visualisation were conducted using the Tidyverse set of pacakges (Wickham, 2017). The machine learning was implemented with the XGBoost, caret and rbayesianoptimisation packages (Chen et al, 2017; Kuhn, 2018; Yan, 2016).

We conducted an extensive descriptive exploration of the data to deepen our understanding of the dataset. Only a subset of the most relevant findings are presented here.

Our key outcome variable for the machine learning analysis was pupil earnings 6-8 years after GCSE. Figure 1 shows the distribution of this variable in the training data.

*Figure 1: Distribution of outcome variable*

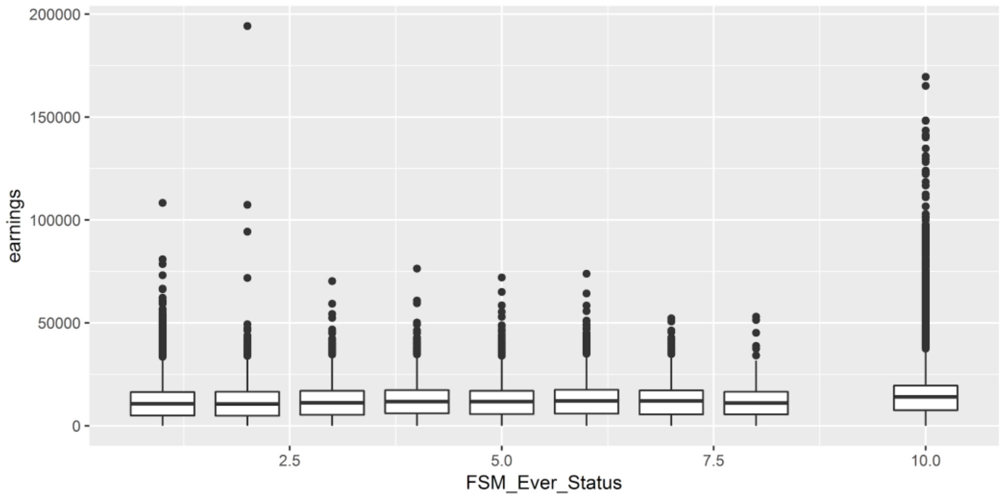


We can see that the variable follows an approximately normal distribution with positive skew. There are also a small number of very extreme outliers at the upper end of the distribution, with one pupil earning nearly £200,000 in their 8th year after completing their GCSEs. At this stage, we won't transform the outcome variable or remove outliers to avoid removing potentially useful information. We will retain all the data for this stage, and assess the model's performance, making adjustments later if necessary.

Next, we explored some key data features relating to social mobility. The department's key economic disadvantage measure is eligibility for free school meals (FSM). FSM is a means-tested benefit given to children from poorer households, and is frequently used in the education research literature as a proxy for disadvantage. This is often used to engineer an 'EverFSM' measure, representing how recently a pupil was eligible for FSM (e.g. the Pupil Premium policy, which grants schools extra funding for each pupil on their roles with EverFSM <= 6). Pupils with an EverFSM value of 0 were eligible for FSM in the year they took their GCSEs; pupils with EverFSM = 1 were most recently eligible in the previous year; EverFSM = 2 were most recently eligible two years ago, and so on.

Figure 2 shows a boxplot of earnings (y axis) against EverFSM (x axis). The box on the far right (EverFSM = 10) is a dummy category representing pupils who had never been eligible for FSM by the time of their GCSEs.

*Figure 2: Correlation of earnings with FSM EverFSM status*



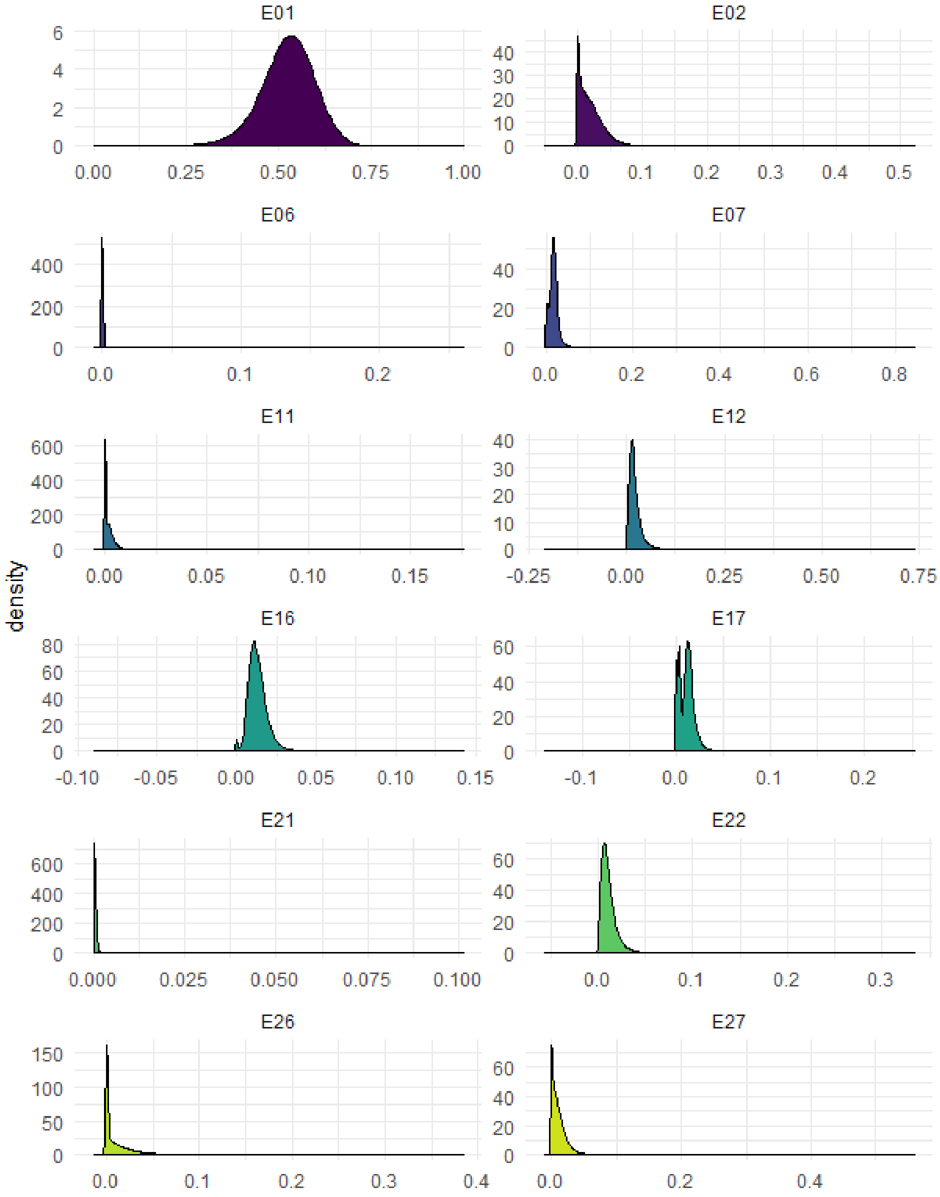
Given the findings in the literature on the strength of the association between parental earnings and those of their children (e.g. Dearden et al, 1995), the association here appears perhaps surprisingly weak. The average earnings of each EverFSM group does appear to increase very slightly as the EverFSM value increases, and is higher for the never-FSM group by around £1500, but there is significant overlap between the groups. The main standout feature of the plot is that the never-FSM group has a much longer tail of high earners.

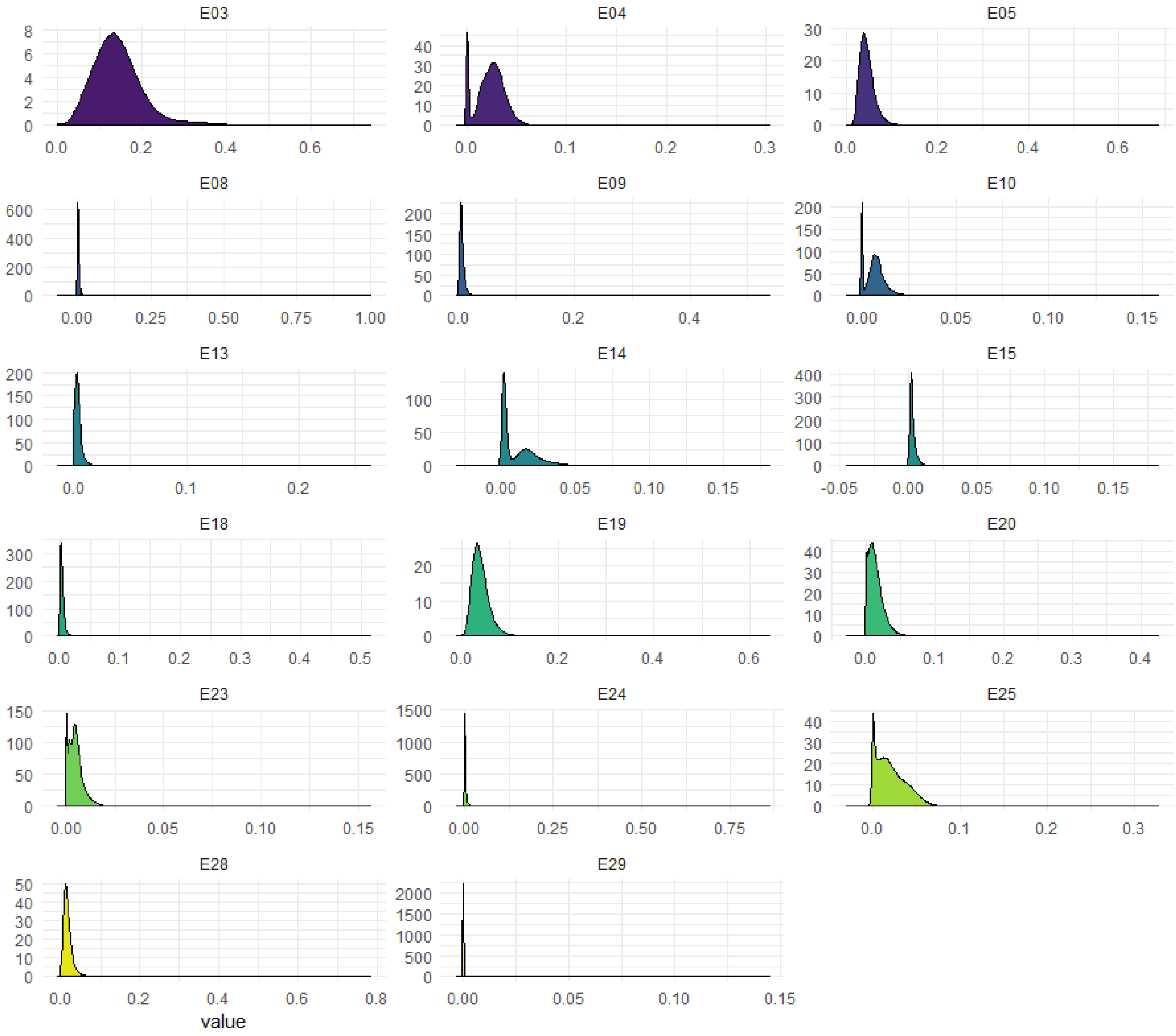
This relatively weak association is most likely because of the fact that there is substantial variation within each of the EverFSM groups in terms of deprivation. A child from a supportive, healthy family that falls just below the earnings threshold for Income Support benefit, for example, will appear equally as deprived as another child from a family living in extreme poverty suffering from severe mental health problems. As such, it will be important to include other measures of deprivation in our model (such as Special Educational Needs (SEN) status, IDACI score, and Children in Need status).

Next, we explored how schools are spending their funding. These variables are included in our analysis as we want to uncover actionable insights that schools can learn from, and spending is one of the areas over which schools (particularly academies) have some control.

Figure 3 shows the distribution of school spending in different categories. Each facet of the chart represents a spending category, and the x-axis shows the proportion of school budget that is spent on that category. The shape of the chart shows how that spending category is distributed across all schools nationally. The spending category codes can be mapped to their descriptions using the index of the latest Consistent Financial Reporting Framework publication.

*Figure 3: School spending distributions by spending category*





Some key findings from this exploration include:

- On average, schools spend more than 50% of their budgets on teaching staff (E01). This is the single largest spending category. No school spends less than 25% of their budget on this category.
- The second largest spend category is Education Support Staff (E03), including teaching assistants. The average spend on this category is around 15% of total expenditure.
- Some other interesting categories that we might want to explore as potential predictors of pupils' later earnings include:
    - Staff development & training (E09)
    - Learning resources (E19)

- ICT learning resources (E20)
- Bought-in professional curriculum services (E27)
- ‘Other staff’, including support staff for pupils with SEN, nurses, medical staff, school counsellors, and so on (E07)

Finally, we conducted an exploration of the geographical spread of earnings and school spending. The notion of ‘places left behind’ has received a great deal of attention politically since the 2016 Brexit referendum, and has been of political interest to education ministers in that time. The hypothesis is that there are disaffected regions of the country (e.g. the West Midlands and the North East) that have been excluded from progress and ‘left behind’ by places like London that have experienced rapid improvement in various educational measures in the last two decades.

Figure 4 shows, to an approximate level of resolution, how the earnings variable in our dataset is distributed across the country. We can see that, as might be expected, alumni of schools in London and the South East have, on average, slightly higher earnings at age 21-23 than those from other parts of the country (with some outlier exceptions). Note that to avoid overplotting, a random sample of 10% of schools are displayed here.

We can see that there does appear to be a differential geographic spread in terms of early-career earnings for the pupils in our dataset. We can also look at how total school spending varies across the country, as shown in Figure 5.

We cannot infer too much from these findings alone, of course: the descriptive earnings picture does not take into account composition or demographic effects, and the average differences between regions masks the fact that there is substantial variation within each area. The spending differential between different areas is also likely accounted for by the fact that school costs are much higher in (for example) London and major cities than in rural areas. Our main conclusion from this exploration is that we should include some area-level and geographic features as predictors in our analysis, to capture any potential geographical element to the relationships we investigate.

*Figure 4: Pupils who went to school in the South East have higher earnings after joining the labour market*

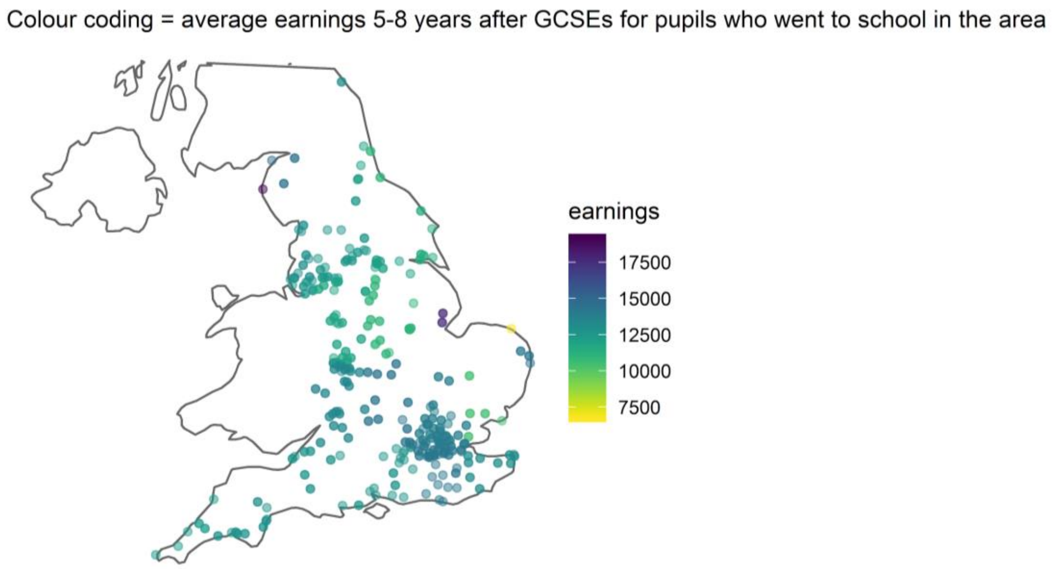
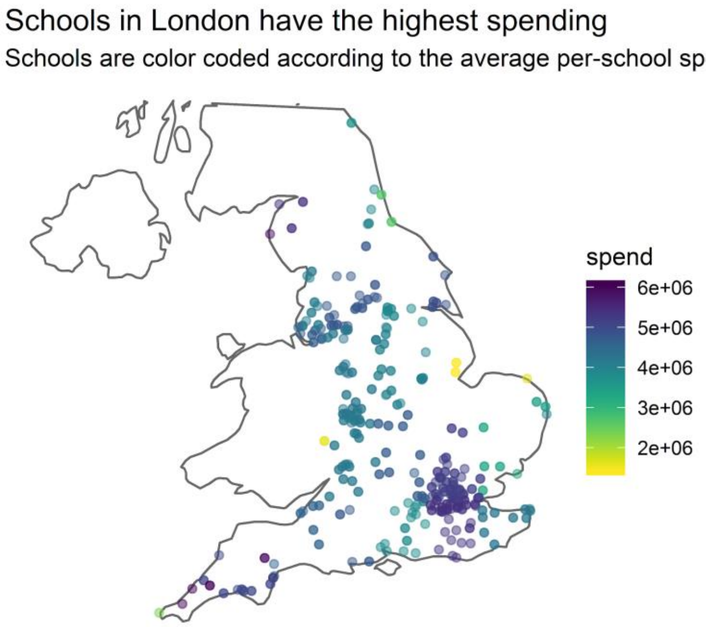

Colour coding = average earnings 5-8 years after GCSEs for pupils who went to school in the area
earnings
17500
15000
12500
10000
7500

*Figure 5: Geographic variation in total school expenditure*

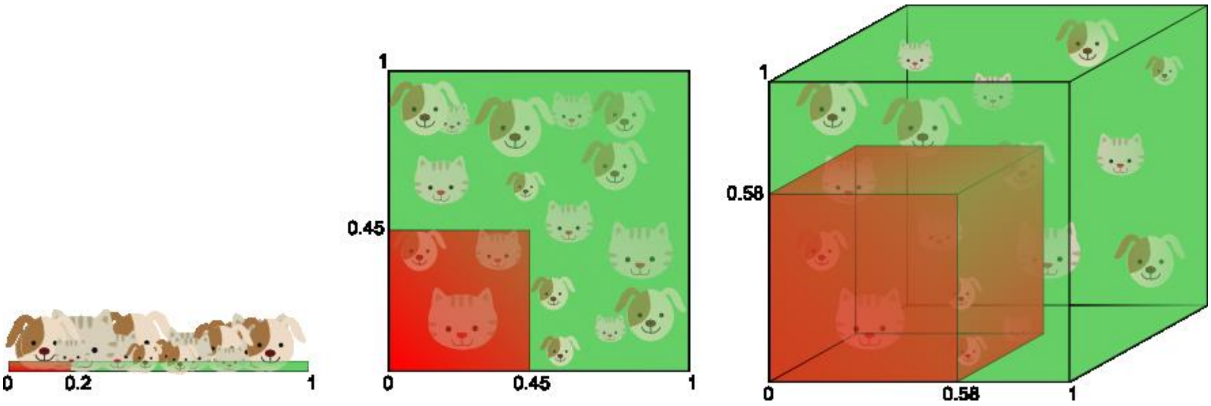

Schools in London have the highest spending
Schools are color coded according to the average per-school spend in their geographic region
spend
6e+06
5e+06
4e+06
3e+06
2e+06

### Dimensionality reduction

After exploring the data descriptively, we next proceeded to prepare our data for the machine learning stage. The aim of this phase was to train a **supervised machine learning model** to predict each individual pupil's earnings as a function of their background characteristics and those of their school.

Machine learning models can suffer from the "curse of dimensionality" (Bellman, 1957). These models all, in some respect, count observations in various regions of the multidimensional feature space. The more dimensions that are present in the data, the fewer observations per region, and so the more prone the model is to overfitting. This is illustrated in Figure 6.

*Figure 6: Illustration of the curse of dimensionality for a classification task*



*Source: MDataGov lecture notes*

One way to minimise this risk is to **reduce the number of dimensions** present in the data. There are a number of techniques that exist for this purpose. In our case, we had a large number of observations (several million) and features (250+), and expected that the dimensions would be non-linearly related to one another. For example, we expect that the association between school spending on SEN provision and Ofsted rating for quality of SEN provision will be moderated by the number of pupils in the school that have SEN conditions.

We therefore required a memory-optimised implementation of some generalised non-linear DR technique. Of the candidates that fell into this category, we chose to use a **deep autoencoding** method. These tend to be fast and effective and can, in theory, approximate any underlying generative function.

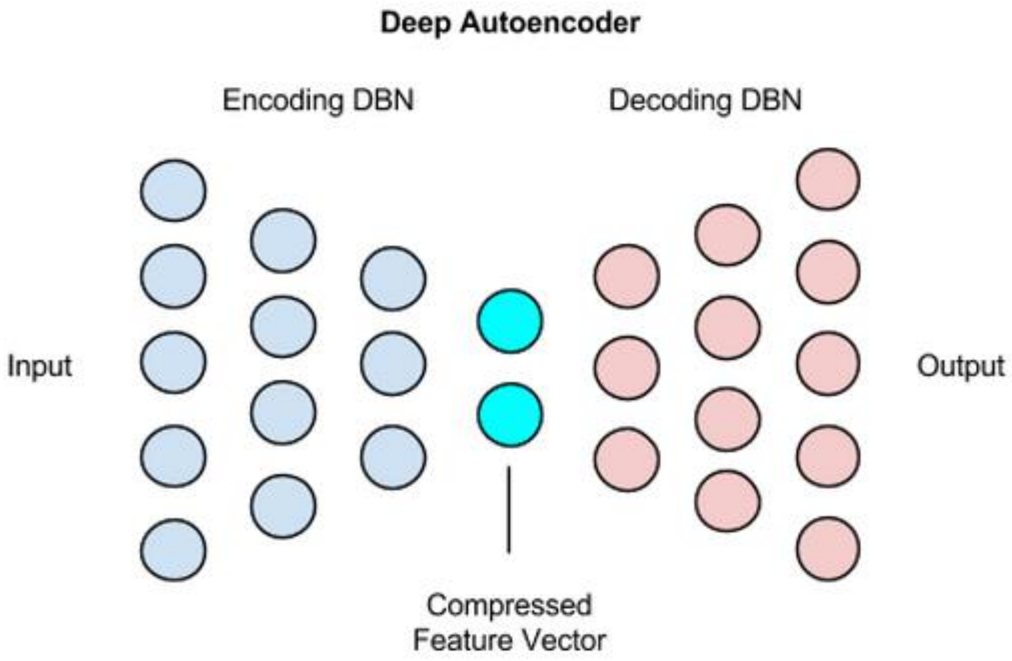
An autoencoder:

> *“Uses a neural network framework to find the most efficient transformation from p dimensions down to whatever you choose. It then finds how well it can reconstruct the original p variables, and keeps tuning until it can optimally recreate the original values as "good" (lowest MSE) as possible”*

(Stack Exchange, 2016, https://datascience.stackexchange.com/a/14296)

Figure 7 shows a diagram of the structure of a typical autoencoder.

*Figure 7: Diagram of the structure of an autoencoding neural network*

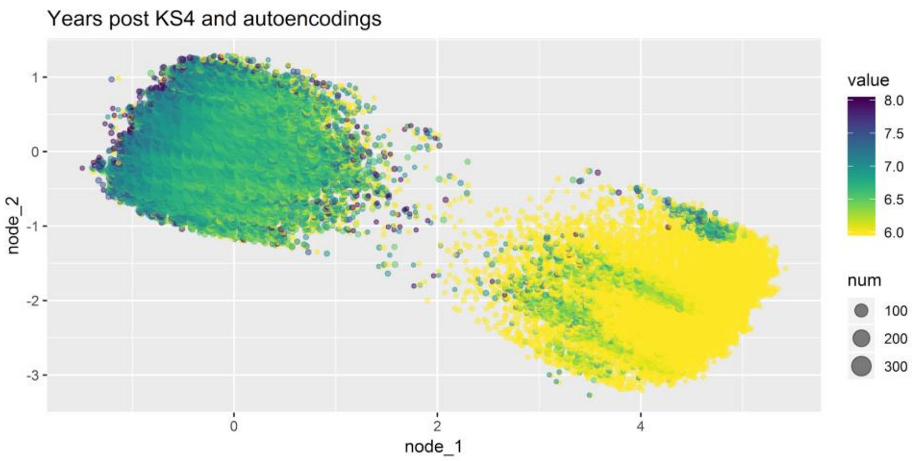


*Source: https://datascience.stackexchange.com/a/14296*

We trained a three-layer autoencoder (AE), with 100 neurons in the first and third layers and 2 in the second. These hyperparameters weren't tuned further, due to time and computational resources available.

We then plotted the cases in the dataset (a random selection, for computational reasons) using the two central features as the x and y axes. The results are shown in Figure 8 below. Note that a random subsample of the data is presented, to avoid overplotting.

*Figure 8: Plot of pupils' central autoencoded features, coloured by years post ks4*



Interestingly, two fairly distinct clusters of pupils emerged. We can see that the right-hand cluster of cases is much more likely to be composed of records relating to earnings data 6 years after GCSE, whereas the left hand cluster is primarily composed of records for later years. This makes sense, as the 6 years post-ks4 group is likely to include many pupils who went on to higher education and only graduated in the middle of that year, and are therefore likely to have different patterns of characteristics. We will account for these nonlinearities by implementing tree-based machine learning models in the predictive section of the analysis.

The 100 features from the first layer were retained as features for training our supervised model, following approaches employed in (e.g.) fraud detection applications.

### Training the model: Reduced feature set

The next stage was to train our supervised model using this reduced set of features as our predictors. In order to do this, we implemented a **Bayesian hyperparameter tuning** process, with **5-fold cross-validation**.

The outline process we followed for training the model was as follows:

1. Remove hold-out data (30% of the hole dataset)
2. Split training data into 5 folds
    a. These folds are each an 80-20 train-test subset of the training data
3. Train a model with an initial set of hyperparameters on each of the 5 folds
4. Test the model on each fold's test set and obtain average error (RMSE)
5. Use a Gaussian process model to estimate the next set of parameter values to test

6. Repeat steps 3 -5 (a total of 20 times)
7. Select model specification with lowest average error
8. Evaluate on hold-out data

**Why did we choose this process?**

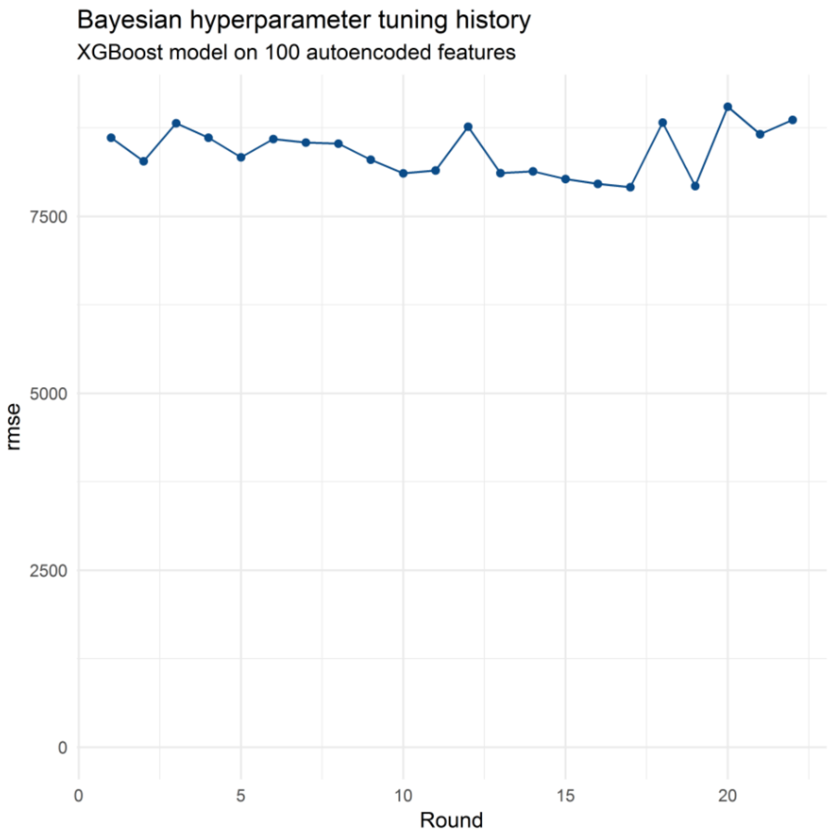
There were several reasons to design our pipeline in this way:

- Hold-out data removed at the start of the process to ensure it could be used to reliably estimate the validity of the model when making out-of-sample predictions
- Model tuning: required as the xgboost framework has a number of hyperparameters that can be varied (e.g. learning rate, maximum tree depth, minimum child weight) and that can have an impact on out of sample predictive power
- 5-fold cross validation was used for model selection as a way of robustly estimating which model specification (i.e. combination of hyperparameters) would be most likely to perform best on hold-out data
- The Bayesian estimation process was implemented as a means to intelligently search for the optimum set of hyperparameter values.  In contrast with other tuning processes (e.g. random search, grid search), the Bayesian process learns from past tuning runs in order to estimate which set of hyperparameters should next be tested. Experimentally, it is often found to result in better predictive power and shorter training times than other common tuning approaches (e.g. Bergstra et al, 2013).

**Testing & evaluation**

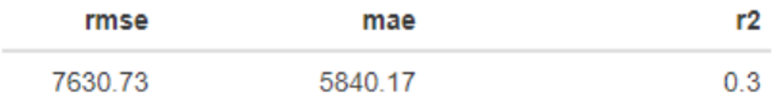
In Figure 9, we can see the model training history. The optimum performance (i.e. lowest error) obtained after 17 iterations.

*Figure 9: Tuning history for the model trained on the reduced feature set*



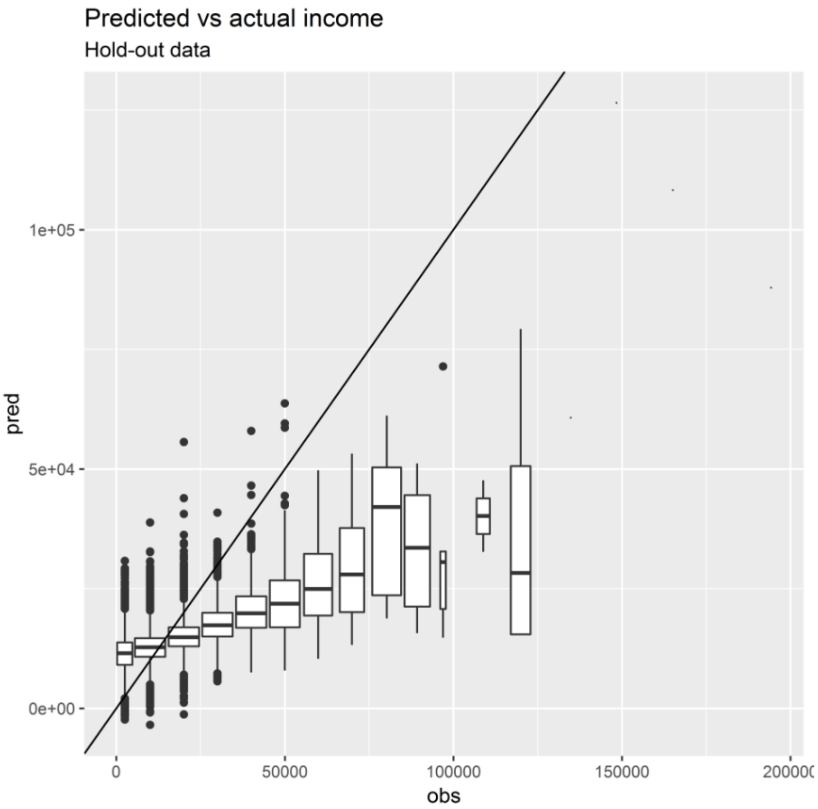
This model’s performance on hold-out data is shown in Table 1 and Figure 10 below.

*Table 1: Reduced feature set XGBoost model performance on hold-out data*



| rmse | mae | r2 |
|---:|---:|---:|
| 7630.73 | 5840.17 | 0.3 |

*Figure 10: Grouped observed values vs predicted values*



## Discussion

We can draw several conclusions from this evaluation of the model:

- The boxplot in Figure 10 shows that the model seems to consistently **under**estimate earnings, particularly for higher earners
- The MAE column in Table 1 shows that, on average, the model's predictions are off by around £5,800. Given that the average earnings for pupils in this dataset is around £20,000 this is a fairly substantial error
- The R square of .3 suggests that the model provides a moderately good fit to the data. This is roughly in line with what we would expect from similar studies in the literature. There is clearly much unexplained variance - which is to be expected, given that we don't hold information on pupil motivation, aspirations, expectations, social capital, or anything that happens to them after KS4

These results also have several implications for the next stage of the analysis:

- The fact that the model consistently underestimates higher earners suggests that we might obtain better results if we were to log-transform our outcome variable. In the current version of the model, a prediction that differs from the observed value by £10,000 is penalised to the same extent regardless of what the observed value is. In other words, being off by £10k is seen as equally as ‘bad’ for someone who actually earns £200k as it is for someone who actually earns £20k. In practice, we probably don’t mind being off by £10k on a £200k salary, but the same error on a £20k salary is very substantial. We can account for this in the next phase by log-transforming the outcome variable before prediction.

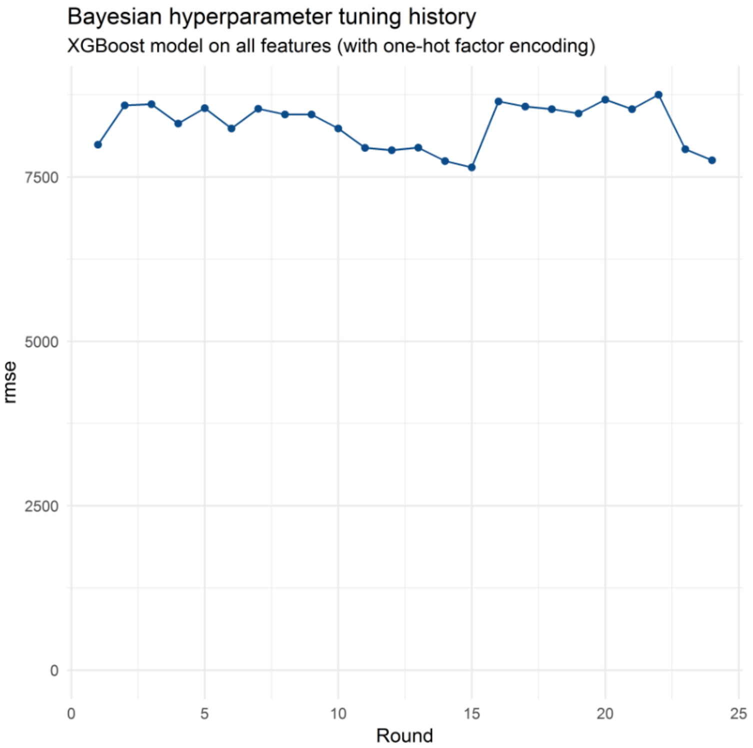
- Training the model on the reduced feature set may have limited its predictive power by reducing the amount of information it can learn from. In addition, it will make it more difficult to identify important variables (e.g. from feature importance plots), as these will be labelled in the model according to the neural network nodes that generated them (not the actual underlying features). We will therefore use the full feature set for the next model training stage and compare its performance with the performance of this model.

### Model training 2: Full feature set

We repeated the training process outlined above using **all** original dimensions[2]. The lowest training error was found after 15 iterations, as shown in Figure 11.

*Figure 11: Training history, XGBoost model trained on all features*

[2] Note that this stage was also conducted with the log-transformed outcome variable, with almost identical results



The boxplot in Figure 12 shows that the model is still underestimating high earners, but to a lesser degree (the box averages are closer to the diagonal line).

*Figure 12: Grouped observed values against predicted values*

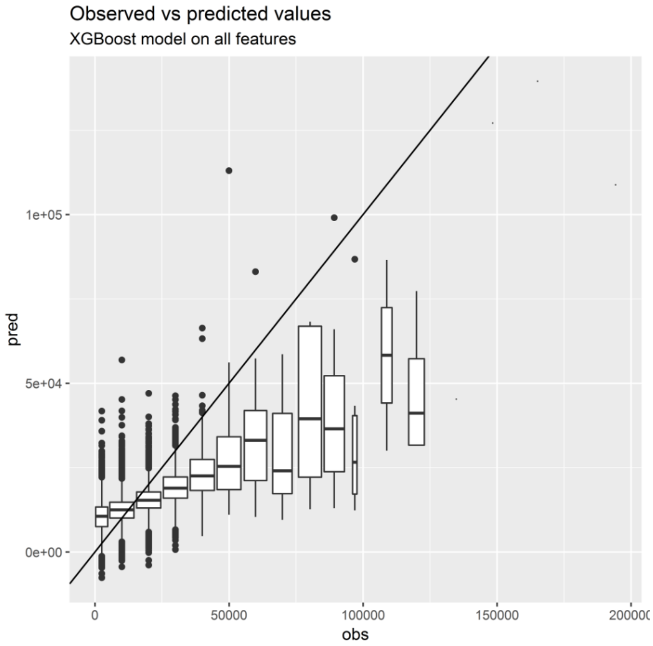
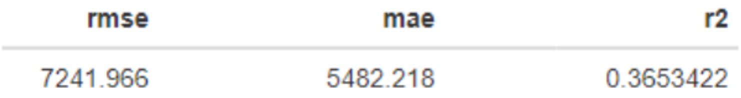


Table 2 shows the error metrics when the model is evaluated on hold-out data. Note that the RMSE and MAE for this model are lower than for the model obtained from the first stage, and the R square value is higher.

*Table 1: Reduced feature set XGBoost model performance on hold-out data*



| rmse | mae | r2 |
|---|---|---|
| 7241.966 | 5482.218 | 0.3653422 |

Given these improved results, we will retain this as our final model.

**Pupil- vs. school-level results**

In addition to these tests, we also evaluated how the model performs when results are aggregated to **school-level.** The reason for doing so is that many of the potential business applications for such a model could be implemented at school level. Examples could include providing greater funding to schools predicted to have low-earning pupil cohorts; hiring careers advisors for these schools; putting local work-experience schemes in place for areas with larger numbers of such schools, and so on.

To perform this evaluation on hold-out data, we took the pupil-level predictions, averaged them by school, and compared them to the actual average earnings of each school’s cohort.

The chart on the left of Figure 13 shows the **pupil** level predicted and actual earnings distribution. The chart on the right shows this at a **school level**. We can see that the overlap is much closer for the school-level predictions.

The performance metrics are also substantially better for the school-level predictions. The MAE column in Table 3 shows that, on average, the model can predict the average earnings of a school’s KS4 cohort to within £900. The R square of 0.84 shows a good fit to the data.

*Figure 13: Pupil and school-level predictions*

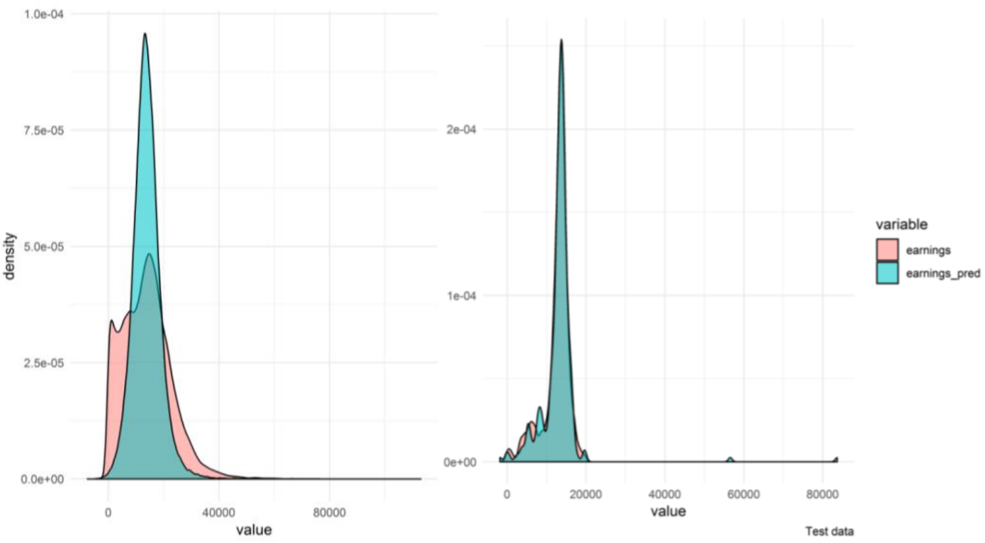


*Table 3: Pupil and school-level predictions*

| | Pupil level | School level |
|---|---|---|
| r2 | 0.37 | 0.84 |
| rmse | 7242 | 2206 |
| mae | 5482 | 895 |
| mae/sd | 0.61 | 0.25 |

Some of this improved performance is, of course, simply a mathematical result of grouping the data together. In averaging over a group, we would always expect the variation to decrease. However, we can see from Table 3 that the results are better at a school level even when accounting for this fact: the MAE as a proportion of total variation in the school-level data (as measured by the standard deviation) is lower for the school-level than the pupil-level predictions.

**Discussion**

This phase of the analysis generated some useful discussion points:

- The higher performance and greater transparency of the non-dimension-reduced model render it more useful than the initial model for influencing the next stage of our analysis. It is interesting to note the lower performance of the dimension-reduced model. Exploring other methods of dimensionality reduction (e.g. clustering, principal component analysis, more finely-tuned autoencoders, etc) could provide a topic for an extended piece of research in itself
- Even if the autoencoded model had performed better, it would have been much more difficult to use for generating a hypothesis to test in Part 2. This is because it would be difficult to identify which features were important for generating the model's predictions - the features would be named after nodes in the neural network (e.g. "Node 1", "Node 2" etc) and not after the underlying variables from which they are created. This provides a good example of the need to balance predictive power with transparency when using machine learning models for practical business purposes.
- The school-level predictions could have many useful applications in a DfE policy context, as many policy interventions are delivered at the school level. In addition to the suggestions above, we could use the model to (for example) identify schools that have high predicted earnings for disadvantaged pupils, and encourage other schools to learn from what they are doing. More analysis would need to be done to investigate the validity of such approaches.
- The fact that the model trained on all features performed better than the dimension-reduced model may also suggest that there are some fine-grained interactions present in the data that influence earnings and that become obscured when the number of dimensions is reduced. Unpicking some of these interactions will be the key focus of Part 2 of this project.

## Part 2: Statistical modelling

In the second stage of the analysis, we build and test several hypotheses about the ways in which DfE can contribute to improving social mobility in England. We did so by:

1. Using the results from the machine learning stage in part 1 to generate data-driven hypotheses about the factors that have the strongest impact on disadvantaged pupils’ future earnings
2. Using robust statistical techniques to test these hypotheses on new data
3. Conducting diagnostic checks and visualisations to verify, validate and interpret the model used for hypothesis testing

Each of these steps is explained in detail below.

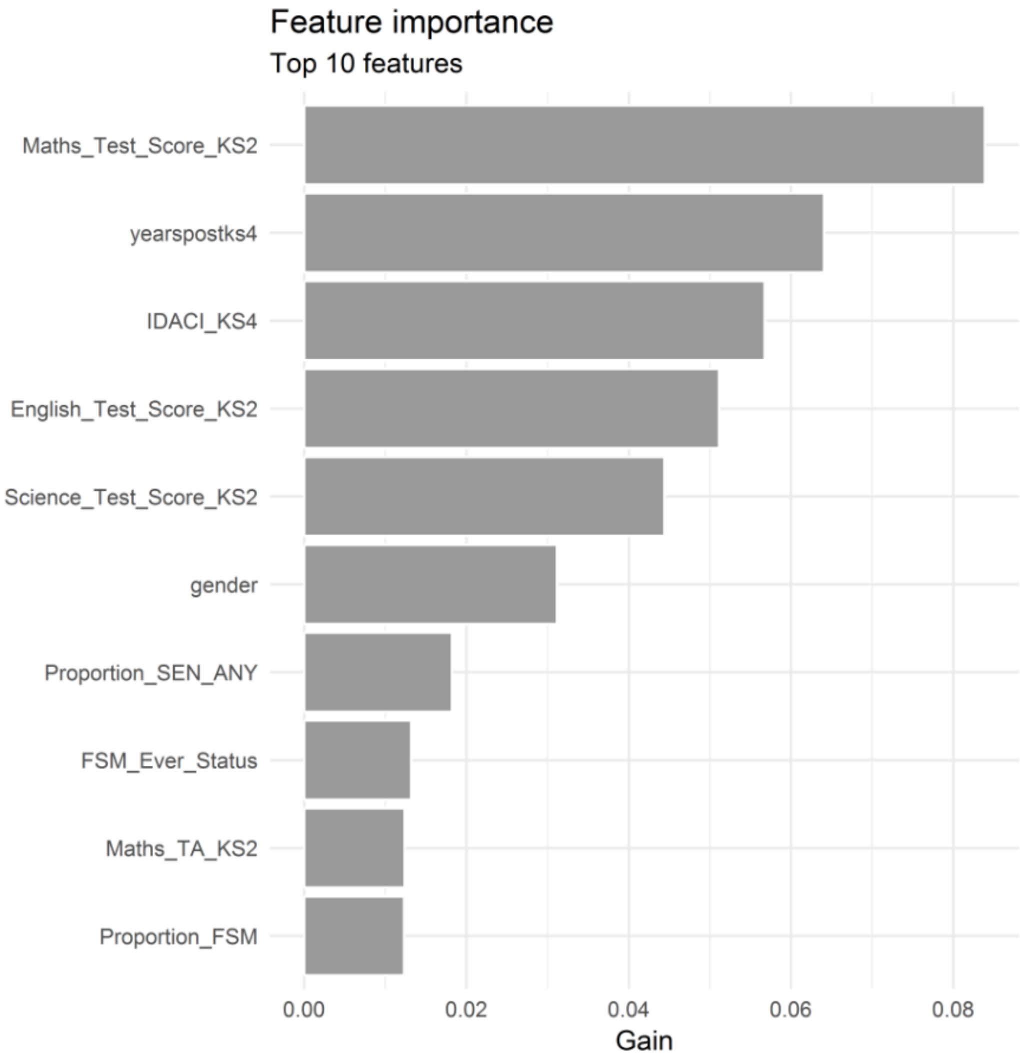
**Hypothesis generation - feature importance**

Our aim in this step was to identify which variables looked most likely to be related to disadvantaged pupils’ future performance in the labour market.

To do so, we pulled out the importance of the features used by the XGBoost model from part 1. ‘Importance’ represents the usefulness of a feature for the model’s **prediction**. More ‘important’ features bring more accuracy to the branches they are on, thus providing the greatest boost to the model’s predictive power (see e.g. *Uğuz, 2011* for a more detailed explanation). The magnitude of a feature’s importance does **not** imply anything directional: if a variable has a high/low importance score, it does not necessarily mean that higher/lower values of that variable are associated with higher/lower values in the response variable.

A plot of the importance for the top ten features is shown in Figure 14.

*Figure 14: Feature importance extract from the XGBoost model, top 10 features*



As we would expect, variables relating to pupil characteristics are the most important. The top three are KS2 performance, years post KS4, and IDACI.

The fact that KS2 maths performance comes out as the most important makes sense. Prior performance is consistently found in the literature to be a very strong predictor of performance at KS4 (Lessof et al, 2018), and KS4 performance is a good predictor of earnings (Hayward & Hunt, 2014). In addition, this prior performance variable may also serve as a proxy for harder-to measure factors such as parental engagement and pupil motivation, both of which have been shown to have a strong impact on KS4 performance in studies of longitudinal survey data (Lessof et al, 2018).

The fact that the ‘years post ks4’ variable is highly important also makes sense. People tend to earn more the longer into their careers they are, so we would expect this to be a useful feature when predicting earnings.

The relatively high importance of IDACI is of particular interest to this investigation. IDACI is the Income Deprivation Affecting Children Index, a composite measure relating to the postcode in which the child lived when they took their GCSE exams. IDACI is strongly correlated with individual and family disadvantage (ref): as IDACI relates to small areas (postcode), it closely reflects the disadvantages faced by individuals living within those areas. As we are concerned with the performance of these disadvantaged pupils, and as the machine learning stage finds that the IDACI disadvantage variable is highly useful when predicting later income, we will retain the IDACI variable for the second stage of the analysis. Given the well-publicised literature on the gender pay gap, it is striking that the model finds IDACI to be almost twice as useful as gender when predicting somebody's later income.

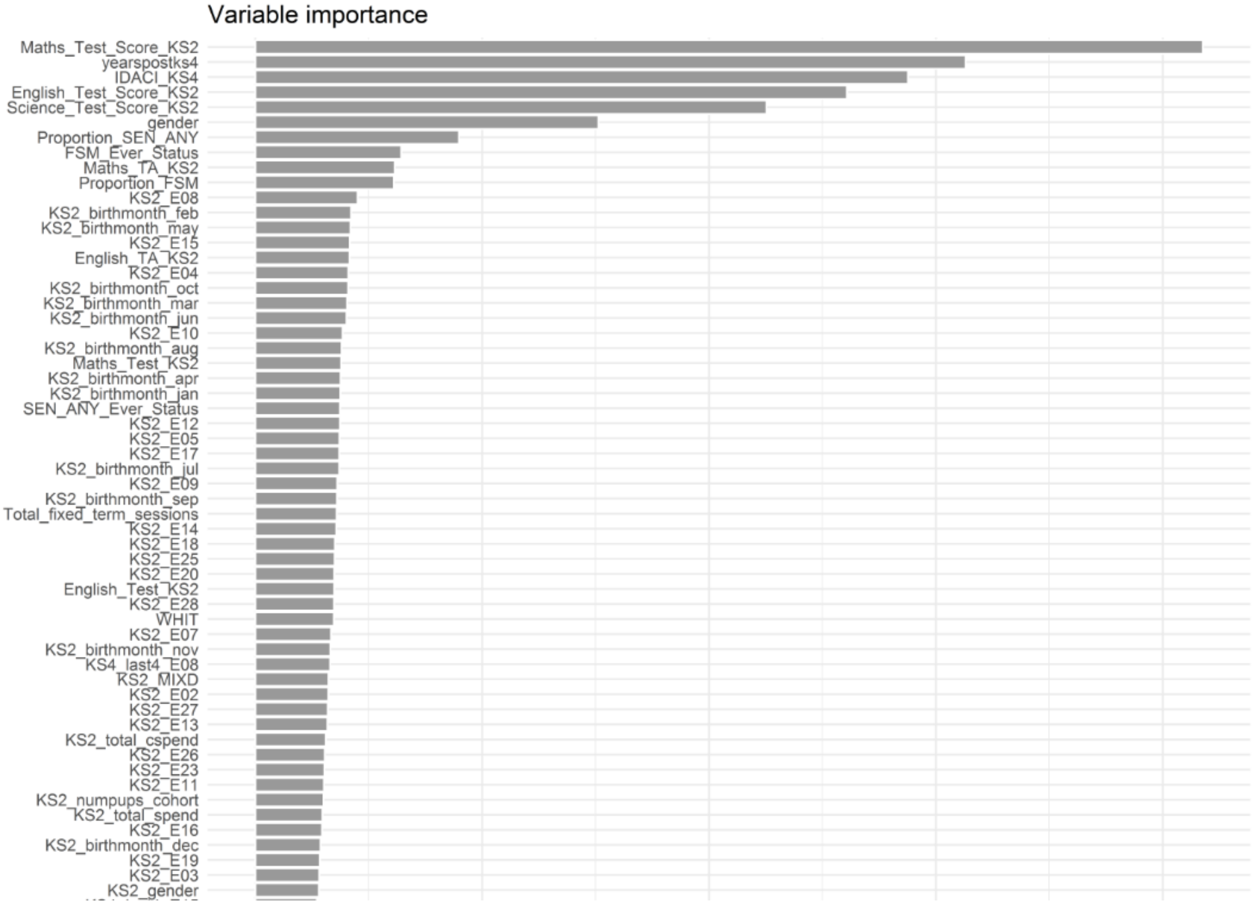
An extended plot of variable importance extracted from the model is shown in Figure 15. Only the top 50 features are shown.

Some other interesting findings from further analysis of the importance scores include:

1. A large number of variables have almost no impact on the model's predictive power. Of the school-level variables, these include: School being a Free School, Community Special School, voluntary controlled school, studio school, community school; gender of entry (mixed/single sex), Ofsted's ratings for KS4 governance value added, effectiveness of safeguarding, and effectiveness of leadership and management; Christian religious character; admissions policy (selective vs comprehensive)
    a. Some of these findings are potentially controversial, and could merit an entire research project in their own right. For example, we find that when predicting a pupil's alter income, knowing whether that pupil went to an academy or not makes no difference whatsoever to the model's prediction. This is a potentially damning finding, given that the government has spent an estimated £750 million to date converting maintained schools to Academies (National Audit Office, 2018).
    b. For the purposes of the current investigation, this finding is useful as it means we can rule these variables out when generating our hypotheses

about which factors *do* improve disadvantaged pupils' labour market performance.

*Figure 15: Feature importance extracted from the XGBoost model*



2. In general, school-level variables are more important at KS2 than KS4. In other words, variables that relate to a child's primary schooling are more useful for the model than variables relating to the child's secondary school. The most important KS2 school variables are:
    a. E08 (indirect employee expenses), E15 (water and sewerage), E04 (premises staff), E10 (supply teacher insurance)
    b. E12 (buildings improvement), E05 (admin staff, including business managers & bursars), E17 (rates), E09 (staff development & training)
    c. Birth months of the KS4 cohort

It is worth re-emphasising that a variable's importance does not imply a positive or negative effect. It simply means the model is made more accurate by including those variables.

Our hypothesis is that spending on E08, E15 and E10 (bullet 2a above) generally have a negative effect on pupil outcomes, acting as proxies for other issues. For example, "indirect employee expenses" includes recruitment costs (which may be indicative of teacher supply issues), travel & subsistence, and compensation. Water & Sewerage spend might indicate schools getting very poor deals on their utility bills, hindering their ability to spend on more useful categories. Supply teacher insurance may also show that the school is struggling to fill vacancies.

The E12, E05, E17 and E09 spend categories (bullet 2b, above) are interesting, and somewhat harder to hypothesise about. Intuitively, it makes most sense that spending on staff development and training (E09) could have a positive impact on pupil outcomes. In addition, schools that spend more on school bursars and finance managers (E05) might be employing better financial managers, and therefore spending the rest of their budget more effectively, resulting in positive outcomes for pupils.

3. The most important variables relating to pupils' secondary schools are:
    a. Number of pupils in the KS4 cohort
    b. E08 (indirect employee expenses), E15 (water), E02 (supply teaching staff) and total Capital spend
    c. Proportion of pupils in the cohort in the Asian ethnicity category
    d. Proportion of pupils in the cohort born in February, April, and August
    e. E18 (other occupation costs - inc. rents, refuse collection, hygiene services etc), E07 ('other staff' - including SEN assistants, liaison officers, careers advisors, youth workers, etc), E22 (administrative supplies), and E05 (admin & clerical staff, including bursars and business managers)

The E07 spend category ('other staff') is potentially interesting here. As it includes spending on assistants for pupils with special educational needs (SEN), we might expect an interaction effect between spending on this category and prevalence of SEN in the cohort. The fact that it also includes careers advisor spend also makes it potentially useful as a predictor of later earnings.

4. Ofsted ratings appear to add very little predictive power on top of the variables discussed above. The most substantial associations are:
    a. 16 to 19 Study Programmes
    b. Personal Development, Behaviour and Welfare
    c. Personal Development, Behaviour and Welfare rating in the previous inspection

The analysis of variable importance in this section has provided some areas for further exploration. Specifically, variables relating to prior performance, gender, ethnicity, SEN, deprivation (IDACI), and school spending on staff and utilities appear to be particularly fruitful avenues for investigation. In the next section, we will use the model to visualise how some of these areas interact with one another. We will then use these visualisations to generate our final hypotheses.

### Hypothesis generation - interaction plots

One of the strengths of tree-based machine learning models is that they are typically very good at taking complex interaction effects into account and using them to make powerful predictions (Schiltz et al., 2018). One of their weaknesses is that it is generally difficult to explore and explain these interactions in an intuitive way. Here, we implement our own version of the partial dependence plot method for visualising specific interactions (see Friedman & Muelman, 2003 for details of the partial dependence plot methodology).

Note that ‘disadvantage’ can take many forms. Here, we will explore effects relating to a number of characteristics known to be associated with relatively poor labour market outcomes. These include gender, ethnicity, SEN, and economic disadvantage.

#### *Gender and KS2 Maths performance*

To create these plots, we take a variable of interest (e.g. Maths KS2 performance) and a grouping variable (e.g. gender). Next, we calculate values corresponding to each decile of this variable in the training data. For each of these values, we take the training data and set the Maths KS2 variable to be equal to that value for all cases. We then use the model to generate predictions for these cases. Finally, we average these new predictions within each level of the grouping variable (e.g. male and female). This procedure is then repeated for each of the other decile values.

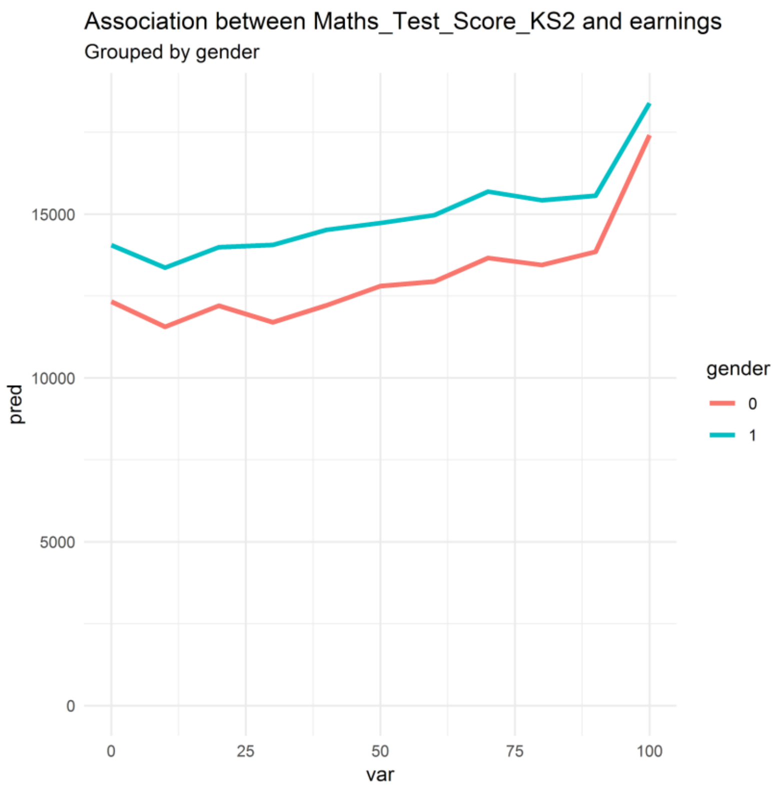
What we end up with is a 10-row dataframe showing what the average earnings prediction *would* be for male and female pupils *if* they all had the same Maths KS2 performance. Each row of the data relates to a different Maths KS2 performance value. Figure 16 shows a plot of this data (note that gender = 1 refers to male pupils, and gender = 0 refers to female pupils). The x axis shows the imputed maths score, and the y axis shows predicted earnings.

There are two important features of this plot:

1. The red line is below the blue line for all values of Maths KS2 score. This shows that the model predicts boys to earn more than girls regardless of Maths score. In other words, boys who do badly at Maths are predicted to earn more than girls who do badly at Maths; boys who do well at Maths are predicted to earn more than girls who do well at Maths, and so on.
2. The lines get closer together as Maths score increases (from left to right). This suggests that doing very well at Maths has a bigger impact on girls' earnings than on boys' earnings.

Taken together, these observations suggest that there may be an interaction effect between gender and Maths performance on later earnings. We will build this into our hypotheses to test in the next stage of analysis. (Note that all hypotheses will be summarised at the end of this section).

*Figure 16: Interaction between Maths KS2 performance and gender*

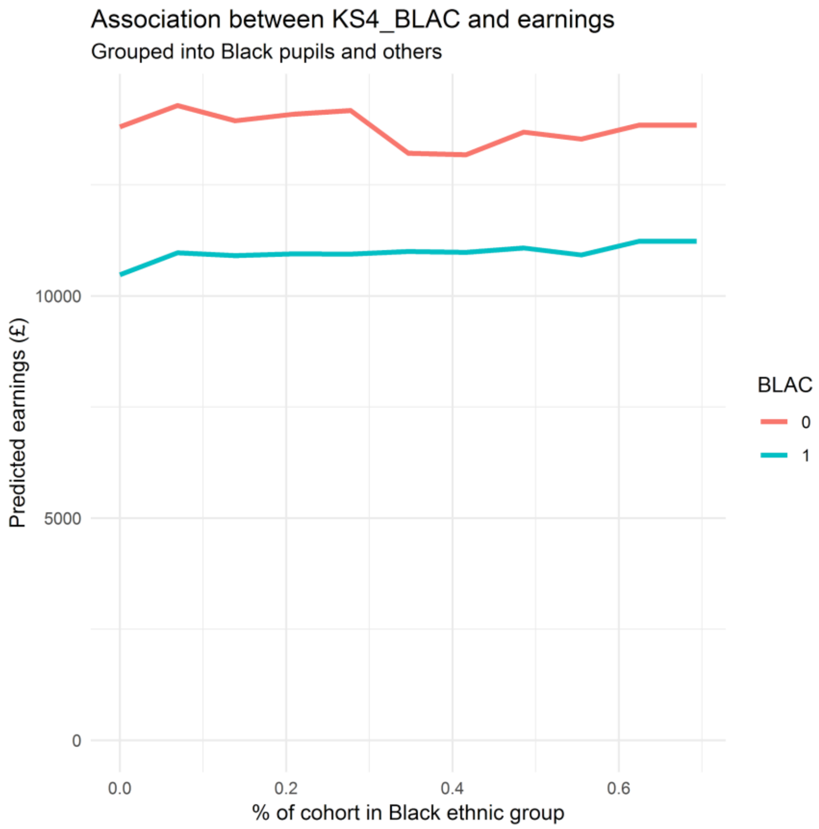


### *Ethnicity*

There are some suggestions in the literature that pupils generally perform worse when they feel alienated from their peer group at school, and better when they feel more included (from Epperson, 1963, onwards). As an important facet of social identity, ethnic grouping can provide a powerful way in which pupils can be made to feel excluded.

To explore this possibility using our data, we investigated whether pupils from various ethnic minorities are predicted to perform better or worse in cohorts where a larger proportion of pupils share their ethnicity, averaging out the effects of other factors. An interaction plot, produced using the same process as for the gender & maths performance visualisation, is shown in figure 17. This plot shows the interaction between Black pupils' earnings and the proportion of pupils in their cohort that share their ethnicity. Similar plots and patterns were observed for other ethnic minorities; for brevity, only one plot is presented here.

*Figure 17: Interaction between Black pupils' later earnings and proportion of KS4 cohort in the Black ethnic minority*



As with the gender/maths plot, we observe:

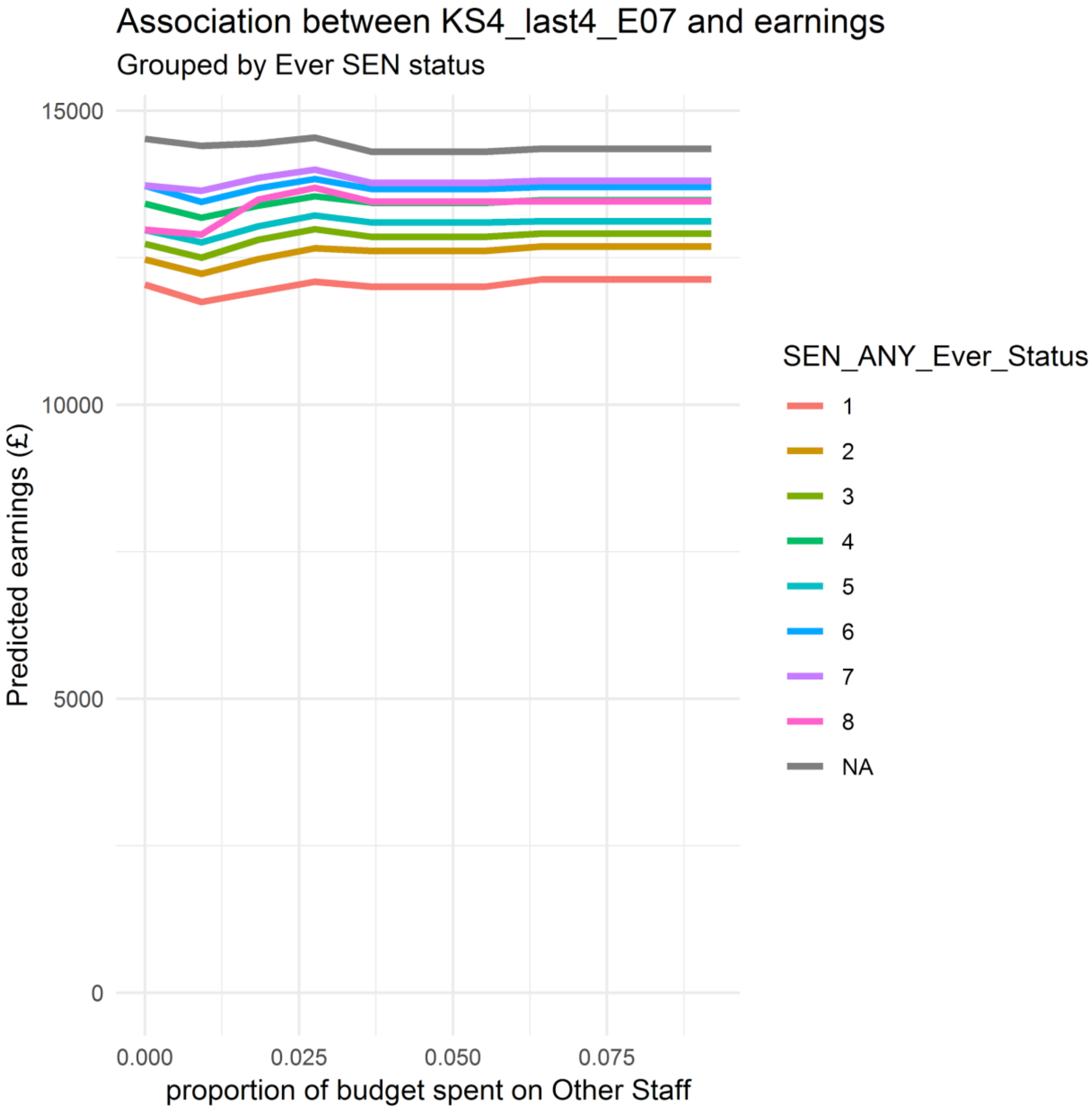
1. Black pupils are predicted to earn less than other pupils (on average) regardless of the ethnicity of their peers
2. Black pupils are predicted to have higher earnings if they attend a school where a larger proportion of the cohort is Black (up to around 70% - the top decile)

Although the increased earnings is relatively small, this finding does support the view that pupils perform better if their ethnicity is less unusual in their school. We will test this view more robustly in the next stage of the analysis, by computing an ethnic diversity measure and including this as a predictor of earnings, interacted with pupil ethnicity. We will use Simpson's diversity index to represent the diversity of the cohort, which specifies the probability that two randomly-selected pupils from a cohort will have different ethnicities (McLaughlin et al, 2016).

*SEN and Other Staff spending*

One of the findings from the variable importance analysis was that both SEN status and secondary school spending on Other Staff (including SEN coordinators and assistants) are relatively important when predicting later earnings. The plot in Figure 18 shows the interaction between these two features predicted by the model.

*Figure 18: Interaction between SEN and secondary school Other Staff spending*



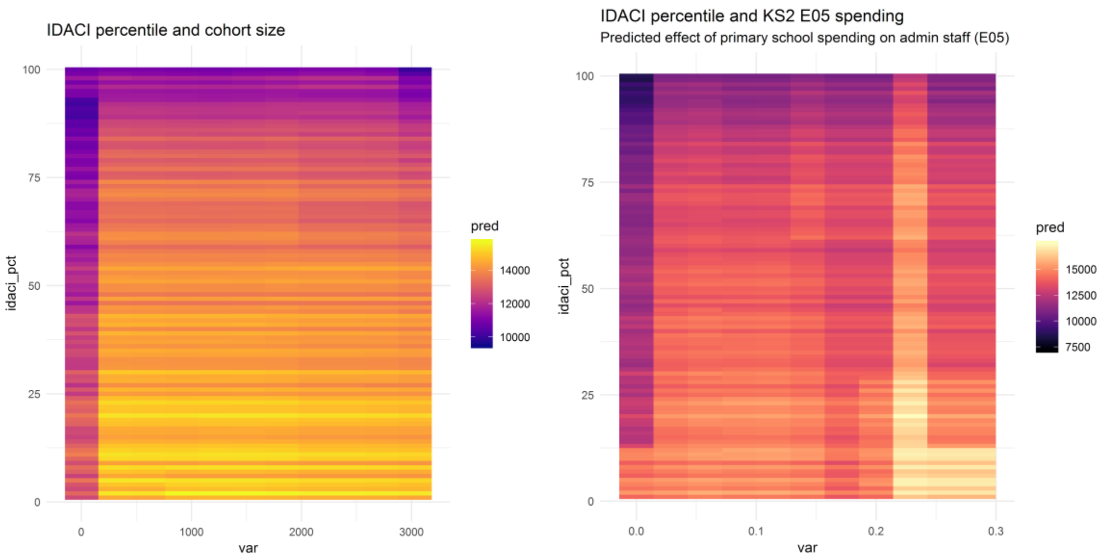
The predictions here are grouped by “SEN Any Ever Status”. This variable relates to how recently before KS4 a pupil was registered as having any special educational need or disability. Those with a value of 1 had a SEN flag in the year they took their GCSE exams; a value of 2 means the pupil had a SEN flag the previous year, and so on. Pupils with an ‘NA’ value had never been registered as SEN.

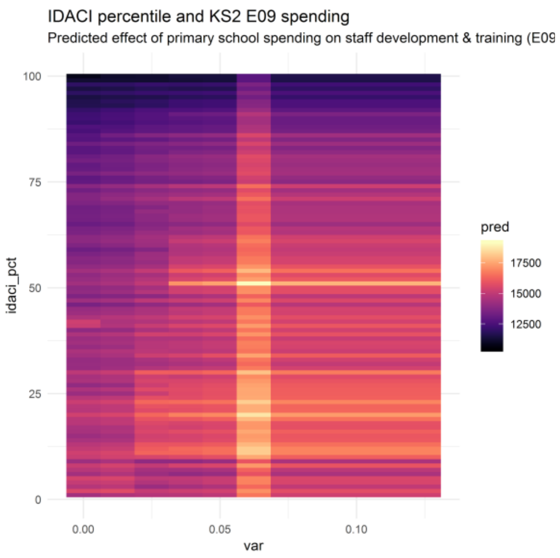
We find that spending on Other Staff (including SEN assistants) has a fairly substantial association with future earnings for pupils who have ever had a SEN flag. The effect is particularly marked for those who had their SEN flag 8 years prior to GCSE (the maximum non-NA value). For example, spending 2.5% of the school budget on Other Staff vs 1.25% has a predicted impact of +£700 in yearly earnings per pupil on average.

*Economic disadvantage*

Next, we produced similar plots showing the average effect of various school spending and cohort characteristic categories, broken down by IDACI score. As there are are many more possible IDACI values than (say) SEN Ever Status categories, the line charts as used above were not appropriate. Instead, the following charts are presented as heatmaps, where lighter colours indicate higher predicted earnings for each combination of IDACI score and values on the variable of interest. Many variables were tested, of which six showed potentially interesting patterns. These are shown in figure 19 below.

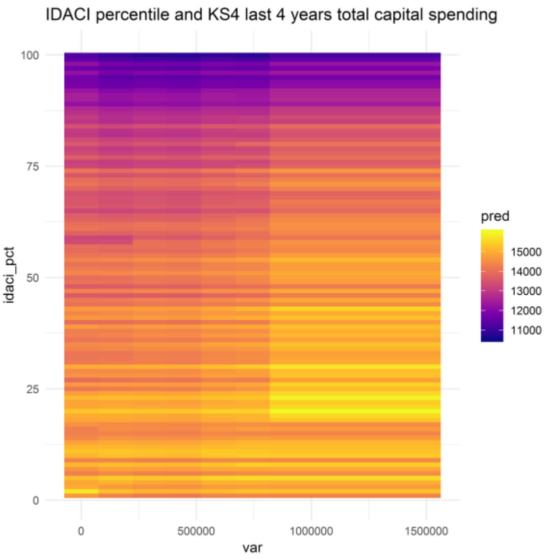
*Figure 19: Interactions between school-level variables and IDACI score*



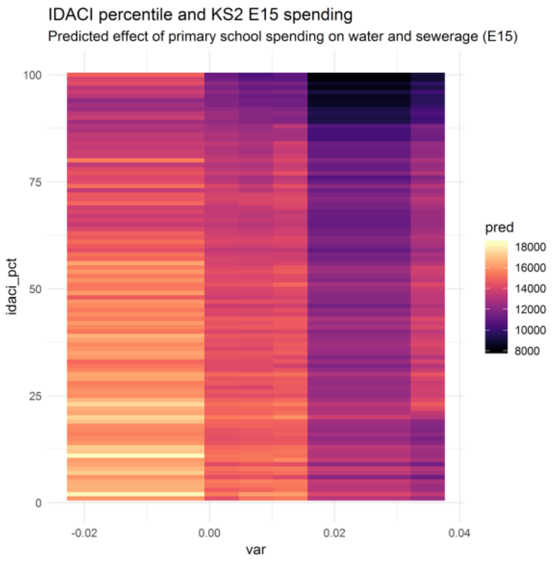


IDACI percentile and KS2 E09 spending

Predicted effect of primary school spending on staff development & training (E09

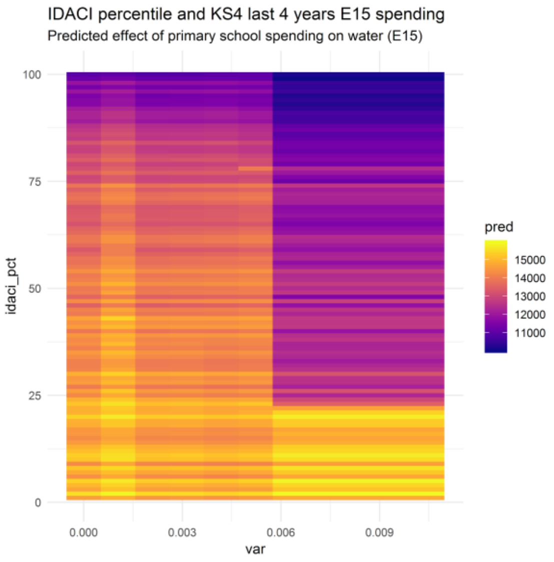


IDACI percentile and KS4 last 4 years total capital spending



IDACI percentile and KS2 E15 spending

Predicted effect of primary school spending on water and sewerage (E15)



IDACI percentile and KS4 last 4 years E15 spending

Predicted effect of primary school spending on water (E15)

We find that:

1. All of the graphs tend to be darker at the top than at the bottom. This shows that pupils living in more disadvantaged areas tend to go on to have lower earnings, in general, than those from less disadvantaged areas.
2. There appears to be a ‘sweet spot’ in terms of KS4 cohort size around the average value. Pupils (especially disadvantaged pupils) in very small and very large schools have lower predicted earnings.
3. Greater proportional spending on admin staff (E05) at KS2 is generally predicted to have a positive effect on later earnings. The effect appears particularly marked for less disadvantaged pupils (note the very light patch in the bottom right)

4. Spending a moderate amount on staff development and training (E09) in primary school appears to be associated with greater pupil earnings, especially for moderately disadvantaged pupils
5. Greater spend on secondary school capital projects is predicted to increase later pupil earnings, especially for those from middle-income areas
6. Greater spending on utility bills (E15) is associated with substantially lower pupil earnings, at both primary and secondary school, especially for disadvantaged pupils.

We took all of the findings from this exploration and used them to generate nine hypotheses to test in the final stage of the analysis. These hypotheses are summarised below.

### Summary

The final hypotheses generated fall into three broad categories:

*Academic effects*

1. Maths performance matters more for girls than boys, in terms of impact on later earnings
2. Maths performance matters more for disadvantaged pupils, in terms of impact on later earnings

*Cohort effects*

3. Ethnic diversity is associated with greater earnings for all pupils, especially disadvantaged minorities
4. Disadvantaged pupils perform better in schools with an average number of pupils in the cohort (i.e. better than in very large and very small cohorts).
    a. Model with a squared term

*Spending effects*

5. Greater secondary school spending on ‘other staff’ is associated with better outcomes for pupils who have ever had a special educational need or disability.
    a. This relationship is moderated by how recently they had this need, with the effect being stronger for pupils affected by SEND longer ago

6. Primary school spending on admin staff (including business managers and bursars) is associated with better performance for all pupils, and the relationship is moderated by disadvantage.
7. There is a nonlinear relationship between primary spending on staff development & training, such that spending a moderate amount (i.e. around 6% of total budget) is associated with better outcomes than spending a large or a small amount. This relationship is moderated by disadvantage.
8. Greater capital spending (e.g. on school improvements) is associated with better long-term earnings
9. Spending more on utility bills (up to around 4% of total budget) is associated with lower future pupil earnings. This effect is particularly strong for disadvantaged pupils.
    a. This effect is observed across both primary and secondary schools

We will test these hypotheses in the next and final stage of the analysis.

**Hypothesis testing**

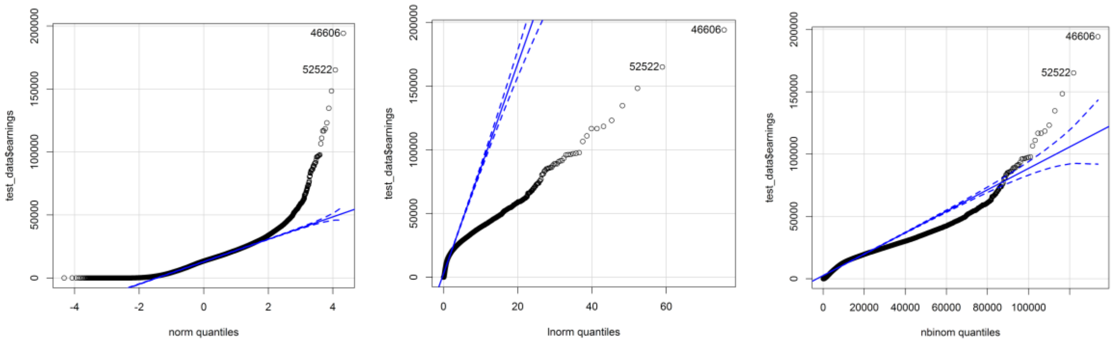
Our aim here was to utilise robust statistical techniques to test these hypotheses on new data, that had not been used for hypothesis generation.

Our data have an inherently hierarchical structure, with observations on the pupil level nested within schools. We know from the literature that peer and compositional effects can have a substantial influence on pupil outcomes (e.g. Levin, 2002), and therefore that we would expect the performance (and later earnings) of pupils from the same schools to be correlated. As such, the assumption of independence required by simple linear models is violated.

We therefore decided to utilise a multi-level (mixed effects) model to address this issue. Mixed models are specifically designed to address this correlation, and will allow us to account for the inherently hierarchical structure of our data.

To decide which family of mixed models to implement, we explored the normality of the response variable (earnings) under several different transformations. These plots are shown in figure 20.

*Figure 20: Non-normality of response variable*



*From left to right: raw (untransformed), lognormal, negative binomial*

The plots clearly show that the response variable does not show a normal distribution in any of the three cases. We therefore chose to use a Penalized Quasi-Likelihood model, which is designed to be robust against non-normality in the response variable, implemented using the MASS package in R (Venables & Ripley, 2002).

The model was specified as follows:

**Model specification**

Response variable:

Pupil earnings

Random intercept effect:

School ID

Fixed effects:

Years post KS4

English Test Score KS2

Science Test Score KS2

Maths Test Score KS2

Gender

KS4 Diversity Index

Ethnicity

SEN Ever Status

IDACI score

Secondary school E07 spending over the 4 years prior to the child's GCSEs

Secondary school E15 spending over the 4 years prior to the child's GCSEs

Secondary school total capital spending over the 4 years prior to the child's GCSEs

Number of pupils in KS4 cohort (squared)

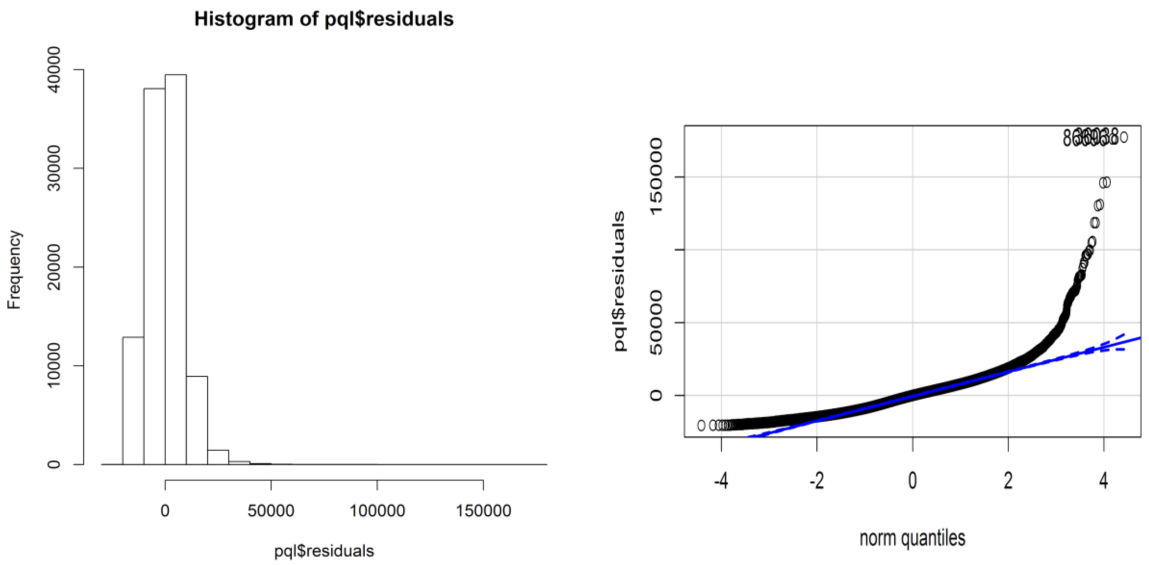
### Evaluation

The model summary output is shown below.

```
## Linear mixed-effects model fit by maximum likelihood
##  Data: test_data
##   AIC BIC logLik
##    NA  NA     NA
##
## Random effects:
##  Formula: ~1 | KS4_LAESTAB
##         (Intercept) Residual
## StdDev:    1133.313 8809.223
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~invwt
## Fixed effects: earnings ~ yearspostks4 + English_Test_Score_KS2 +
Science_Test_Score_KS2 +      Maths_Test_Score_KS2 + gender +
gender:Maths_Test_Score_KS2 +      IDACI_KS4:Maths_Test_Score_KS2 +
KS4_diversity_index + AOEG +      ASIA + BLAC + CHIN + MIXD + UNCL +
KS4_diversity_index:AOEG +      KS4_diversity_index:ASIA +
KS4_diversity_index:BLAC + KS4_diversity_index:CHIN +
KS4_diversity_index:MIXD + KS4_diversity_index:UNCL + SEN_ANY_Ever_Status +
KS4_last4_E07 + SEN_ANY_Ever_Status:KS4_last4_E07 + IDACI_KS4 +      KS2_E15 +
KS4_last4_E15 + KS2_E05 + KS2_E09 + I(KS4_numpups_cohort^2) +
```

Several validation checks were performed on the model. Figure 21 shows that the model’s residuals appear approximately normally distributed, albeit with somewhat high kurtosis and positive skew.

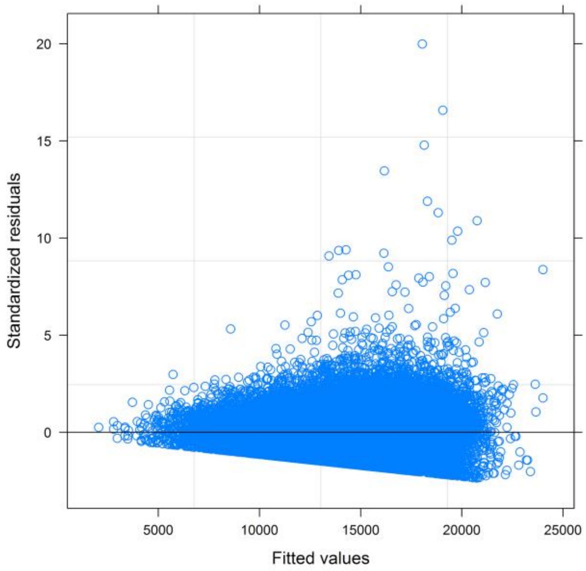
*Figure 21: Histogram and Q-Q plot of PQL model residuals*



The mean of the residuals is 34.24, which in the context of earnings data is very close to zero (0.4% of one standard deviation). This suggests that the model does not suffer from substantial bias.

Figure 22 shows a plot of fitted values against the model's residuals.

*Figure 22: Fitted values against standardised residuals*



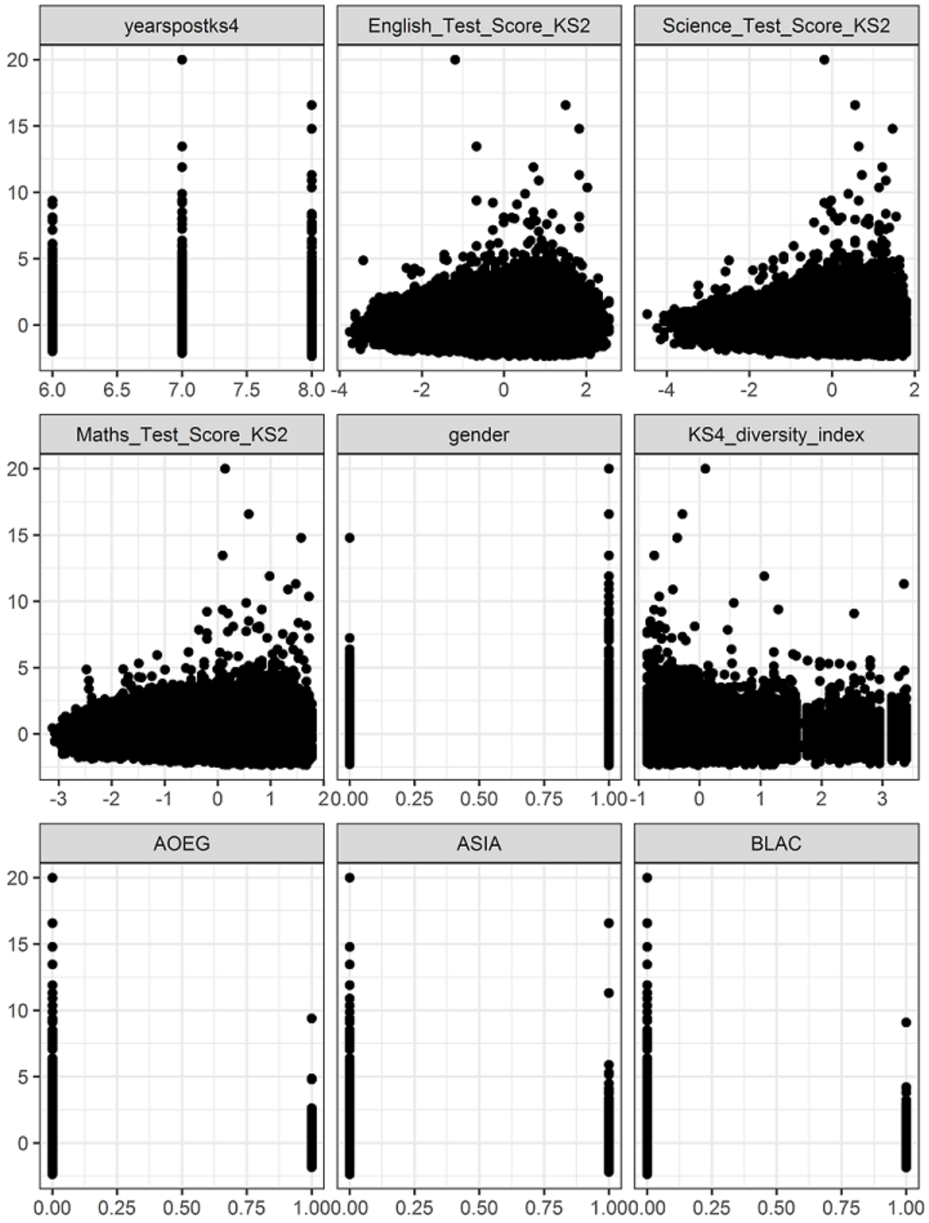
This plot shows that there appears to be fairly substantial inconsistency in the variance of the residuals across the fitted range (heteroskedasticity). In addition, there is an unusual sharp diagonal line in the pattern of residuals along the bottom edge of the plot. This is likely to be due to the truncated nature of the data (i.e. pupil earnings cannot be below zero) and so should not by itself be a major cause for concern.

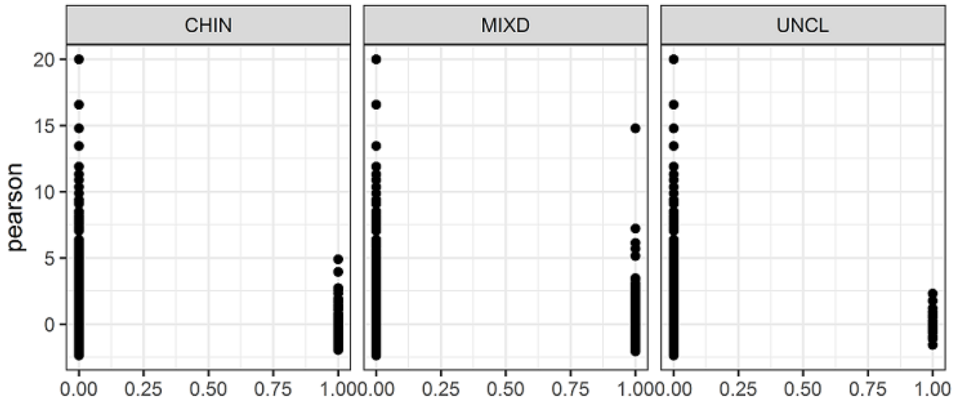
The heteroskedasticity may also be the result of the nature of the response variable. We see that residuals tend to be smaller for lower fitted values. In other words, the model is more accurate when it predicts a pupil's earnings to be lower rather than higher. In the context of earnings data, this might not be too problematic: a residual of £1000 on a predicted salary of £20,000 is much less significant than it would be on a predicted salary of £5,000.

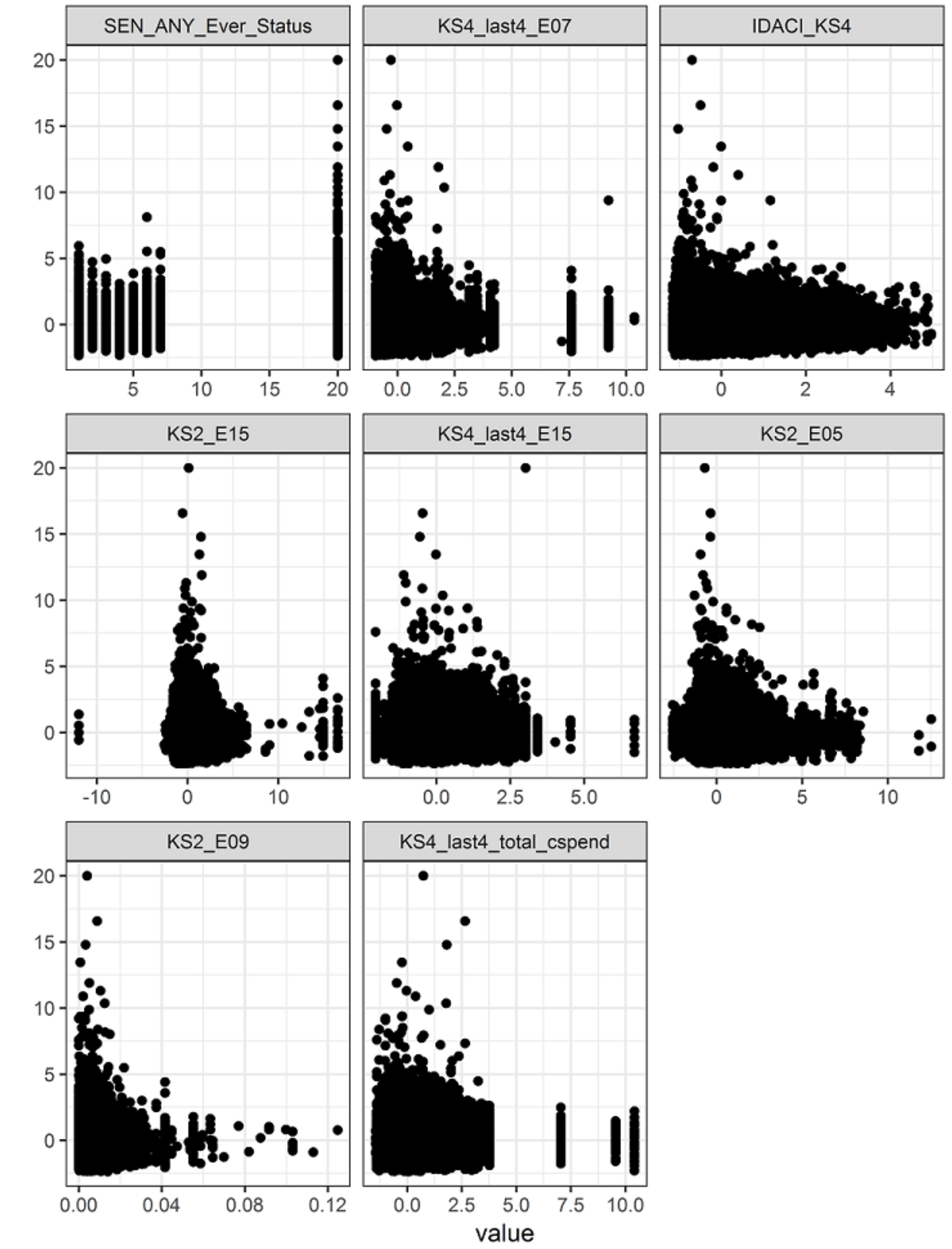
Future work on this data could explore other means of tackling this potential issue (such as through other transformations of the response variable, or use of models that are designed to deal with truncated response variables).

We also examined the model for linearity in the residuals by values of the model's predictors. Figure 23 shows that none of these appears to show cause for significant concern.

*Figure 23: Linearity in residuals by predictors*




CHIN
MIXD
UNCL
pearson
20
15
10
5
0
0.00
0.25
0.50
0.75
1.00


SEN_ANY_Ever_Status
KS4_last4_E07
IDACI_KS4
KS2_E15
KS4_last4_E15
KS2_E05
KS2_E09
KS4_last4_total_cspend
20
15
10
5
0
-10
0.0
2.5
5.0
7.5
10.0
2
4
0.00
0.04
0.08
0.12
value

**Results & discussion**

The full output from the model is shown below.

| | Value | Std.Error | DF | t-value | p-value |
|---|---|---|---|---|---|
| (Intercept) | -3246 | 387 | 50358 | -8.39 | 0.000 |
| yearspostks4 | 2256 | 53 | 50358 | 42.50 | 0.000 |
| English_Test_Score_KS2 | -287 | 67 | 50358 | -4.27 | 0.000 |
| Science_Test_Score_KS2 | 177 | 70 | 50358 | 2.53 | 0.012 |
| Maths_Test_Score_KS2 | 1137 | 81 | 50358 | 14.03 | 0.000 |
| gender | 2231 | 86 | 50358 | 25.96 | 0.000 |
| KS4_diversity_index | 440 | 93 | 50358 | 4.74 | 0.000 |
| AOEG | -2641 | 823 | 50358 | -3.21 | 0.001 |
| ASIA | -820 | 310 | 50358 | -2.65 | 0.008 |
| BLAC | -2829 | 492 | 50358 | -5.75 | 0.000 |
| CHIN | 559 | 910 | 50358 | 0.61 | 0.539 |
| MIXD | -1420 | 309 | 50358 | -4.60 | 0.000 |
| UNCL | -3322 | 2397 | 50358 | -1.39 | 0.166 |
| SEN_ANY_Ever_Status | 77 | 6 | 50358 | 12.39 | 0.000 |
| KS4_last4_E07 | -34 | 114 | 50358 | -0.30 | 0.765 |
| IDACI_KS4 | -461 | 64 | 50358 | -7.19 | 0.000 |

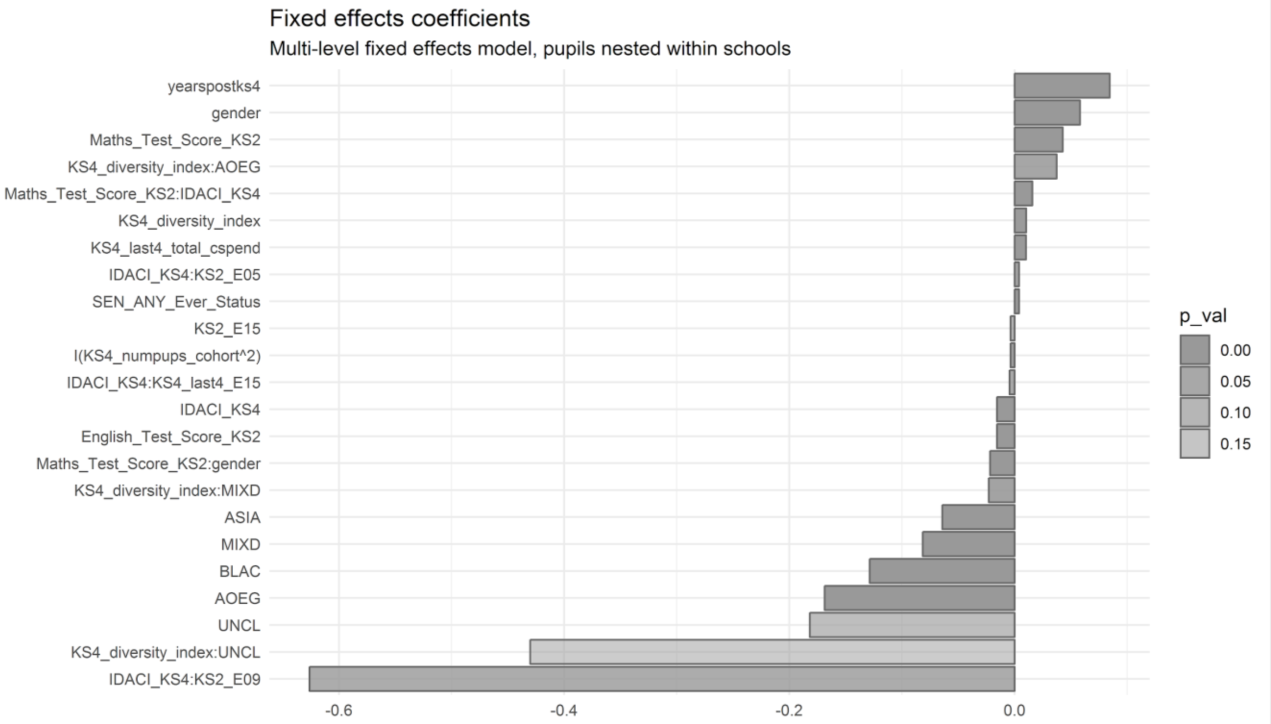
|  |  |  |  |  |  |
|---|---|---|---|---|---|
| KS2_E15 | -17 | 44 | 50358 | -0.39 | 0.695 |
| KS4_last4_E15 | -130 | 82 | 50358 | -1.59 | 0.112 |
| KS2_E05 | 55 | 46 | 50358 | 1.19 | 0.232 |
| KS2_E09 | 2712 | 8381 | 50358 | 0.32 | 0.746 |
| I(KS4_numpups_cohort^2) | -112 | 33 | 50358 | -3.36 | 0.001 |
| KS4_last4_total_cspend | 244 | 68 | 50358 | 3.60 | 0.000 |
| Maths_Test_Score_KS2:gender | -394 | 81 | 50358 | -4.87 | 0.000 |
| Maths_Test_Score_KS2:IDACI_KS4 | 159 | 41 | 50358 | 3.90 | 0.000 |
| KS4_diversity_index:AOEG | 626 | 373 | 50358 | 1.68 | 0.094 |
| KS4_diversity_index:ASIA | -164 | 183 | 50358 | -0.89 | 0.371 |
| KS4_diversity_index:BLAC | -11 | 231 | 50358 | -0.05 | 0.963 |
| KS4_diversity_index:CHIN | -598 | 586 | 50358 | -1.02 | 0.307 |
| KS4_diversity_index:MIXD | -435 | 219 | 50358 | -1.99 | 0.047 |
| KS4_diversity_index:UNCL | -8473 | 6377 | 50358 | -1.33 | 0.184 |
| SEN_ANY_Ever_Status:KS4_last4_E07 | 1 | 5 | 50358 | 0.16 | 0.873 |
| IDACI_KS4:KS2_E09 | -9697 | 6853 | 50358 | -1.42 | 0.157 |
| IDACI_KS4:KS2_E15 | -61 | 43 | 50358 | -1.44 | 0.151 |
| IDACI_KS4:KS4_last4_E15 | -31 | 48 | 50358 | -0.64 | 0.519 |
| IDACI_KS4:KS2_E05 | 48 | 39 | 50358 | 1.23 | 0.220 |

| | | | | | |
|---|---|---|---|---|---|
| I(KS4_numpups_cohort^2):I(IDACI_KS4^2) | 2 | 22 | 50358 | 0.11 | 0.913 |

A plot of the fixed effects values, filtered to those with relatively low *p*-values (p < 0.15) is shown in figure 24. Note that continuous variables are scaled such that the coefficients are expressed in terms of standard deviations in the response variable.

The implications of these results for our hypotheses are discussed below.

*Figure 24: Fixed effects coefficients*



*Hypothesis 1: Maths performance matters more for girls than boys, in terms of impact on later earnings*

Our results show that:

- The coefficient for gender (i.e. being male vs female) is £2,231 (std. erro = £86, $p < 0.0001$)
- An increase in Maths KS2 score by one standard deviation for **girls** is associated with an increase in earnings of £1,137 (std. error = 81, $p < 0.0001$)
- The same increase for **boys** is associated with an increase in earnings of £743 (std. error = 81, $p < 0.0001$). This is shown by the coefficient for the interaction term between gender and KS2 Maths performance (-£394).

These results suggest that we can reject the null hypothesis that the coefficient for the interaction between gender and maths score is equal to zero. We can conclude that maths performance is associated with greater financial returns for girls than for boys.

One possible explanation for this result might be related to gender stereotyping and social norms. We know that mathematical and scientific subjects are generally seen as more male than female in the UK (e.g. Codiroli McMaster, 2017), and that boys are more likely to continue studying maths to higher levels than girls with similar prior mathematical abilities. In addition, there is a well-established body of literature indicating that performance in STEM subjects and achievement of STEM degrees is associated with greater labour market returns than other subjects (Altonji et al, 2012). It seems logical, therefore, that are findings are reflective of a world in which girls need to have very high levels of interest in and aptitude for maths in order to overcome social norms and go on to study it further, and thus gain access to higher paying jobs. Boys, by contrast, are more expected to study mathematical/STEM subjects, and so performance at KS2 matters less in determining whether they go on to these kinds of careers. This would explain the interaction effect observed in this study.

*Hypothesis 2: Maths performance matters more for disadvantaged pupils than non-disadvantaged pupils*

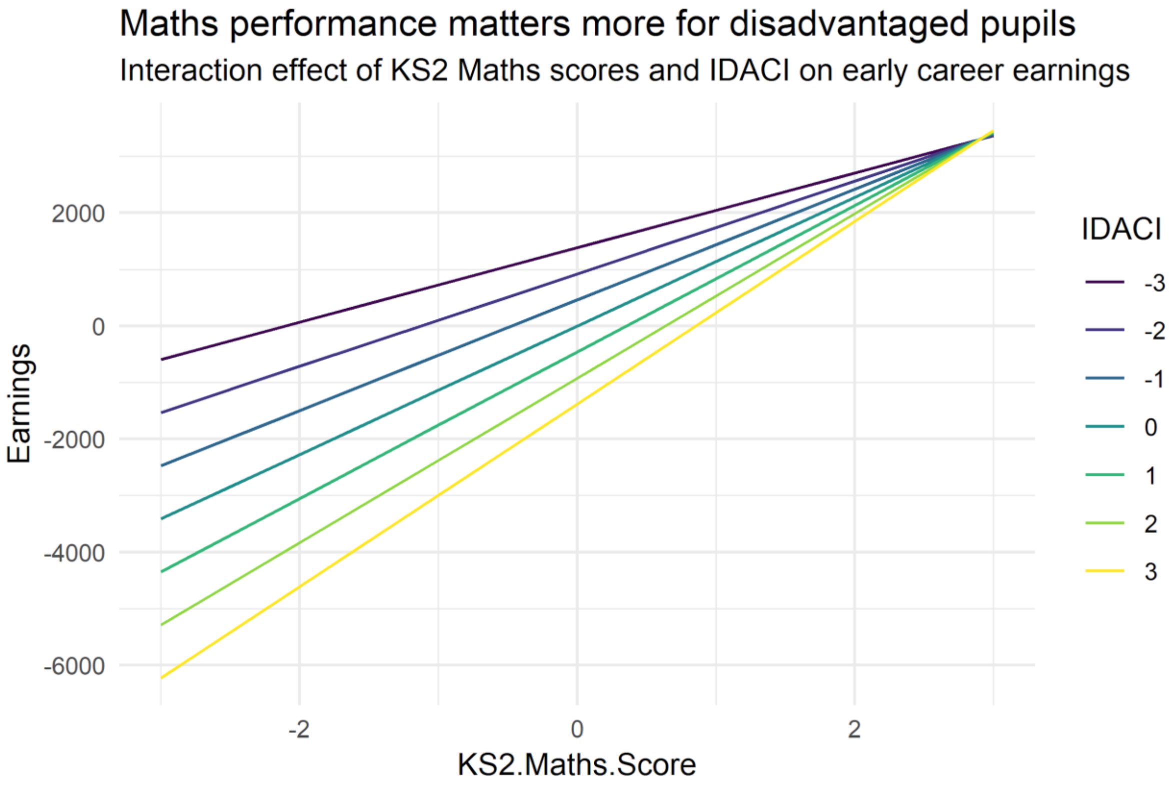
We find that:

- There is a substantial, statistically significant interaction between IDACI and KS2 Maths Score that affects pupils' future earnings (coeff = 159, std. err = 41, $p < 0.0001$)
- Figure 25 provides a striking visualisation of this interaction. We see that, for pupils who do very poorly in Maths at KS2, IDACI has a marked association with earnings, with those living in the most affluent areas going on to earn nearly

£6000 more per year on average than those leaving in the least affluent areas. In stark contrast, for those who do very well in Maths at KS2, the effect of living in a poor area is almost entirely eliminated.

These findings indicate that we can reject the null hypothesis and conclude that both IDACI and KS2 Maths performance are associated with future earnings, and that Maths performance appears to be particularly important for disadvantaged pupils.

*Figure 25: Interaction between KS2 Maths score and IDACI*



We can also provide a solid theoretical interpretation for this finding. Other studies have found that more affluent pupils tend to be “shielded” from the effects of poor schooling and poor educational performance in various ways. For example, children born into wealthy families who do not achieve higher education qualifications are more likely to go on to relatively well-paid jobs than similarly-qualified children born into poorer families (Social Mobility Commission, 2019). There are a host of possible reasons for this, including increased social capital, network effects, social norms, and discrimination in the labour market.

In addition, it is frequently suggested that children from wealthier families who perform poorly at KS2 are much more likely to ‘catch up’ than pupils from poorer families. The wealthy families are generally more able to provide private tuition, separate quiet study spaces, learning resources, and time to focus on studying. This means that those pupils are more likely to be able to recover from their earlier poor performance, go on to achieve better grades later in their school careers, and thus go on to better-paid jobs than their more disadvantaged peers.

Finally, social norms and expectations may also be a contributing factor. The class system in the UK generally sees academic aptitude as associated with the wealthier middle classes, and this bias may be held unconsciously by teachers and pupils in the education system. As such, if a child from a more affluent family achieves poor KS2 results, it might be that they, their family and their teachers are more likely to see these as anomalous or uncharacteristic results. If a child from a more disadvantaged background gets the same results, it may be that this is more likely to be seen as reflective of the child’s inherent aptitude, motivation, or personal character. This could, in theory, lead to the wealthier child being less discouraged by their poor prior performance, and treated differently by teachers and family alike during the rest of their academic career. This kind of self-fulfilling prophecy effect is well-documented in the literature on social mobility, and may help to explain the effects we find here (e.g. Rist, 1970; Willard & Madon, 2016, amongst others).

*Hypothesis 3: Ethnic diversity is associated with greater earnings for all pupils, especially disadvantaged minorities*

We find that:

- Greater diversity is associated with higher earnings for White pupils: a one standard deviation increase in diversity index is associated with an increase in yearly earnings of £440 (std. error = 93, $p < 0.0001$)
- The net effect of diversity on earnings is more or less zero for pupils in the Mixed ethnic group. The base effect of +£440 is offset by the -£435 coefficient for the Mixed group interaction term (std. error = 219, $p < 0.05$)

- For the other ethnic groupings, a larger sample would be needed to determine whether the effect of diversity on earnings is different to that observed for White pupils (all *p* values >> 0.05)
- We therefore cannot reject the null hypothesis with respect to the *interaction* terms with this data. We do not have sufficient results to conclude that the coefficients for the interaction terms between diversity index and ethnicity are nonzero in the population.

These findings suggest that can reject the null hypothesis and conclude that greater diversity in school cohorts is associated with better earnings later in life. For White pupils, the increase is fairly substantial: an average of £440 per pupil per year for each standard deviation increase (around 17 percentage points) in diversity index. The fact that the interaction terms are almost all non-significant means that this increase may or may not be different for other ethnic groups.

These findings make sense intuitively. We would expect pupils to perform less well if they are ostracised (or even bullied or abused) as a result of their ethnicity, and we would expect this kind of treatment to be more common in schools where ethnic diversity is low (as ethnic minorities would be more obviously the “odd ones out”).

Another potential reason for this observation is that schools in London perform particularly well, and are also more likely to have ethnically diverse pupil cohorts (Burgess, 2014). The literature is still unclear on the extent to which this greater performance is caused by ethnic diversity, as opposed to other factors (e.g. the increased funding these schools received under the Blair government as part of the London Challenge). Recent work by the FFT Education DataLab (2019) replicated Burgess’s methodology and found a moderately substantial London effect (around one quarter of a GCSE grade per pupil) even after controlling for ethnicity and demographic factors. We therefore cannot tell whether our results are indicative of a positive causal relationship between ethnic diversity and future pupil earnings, or whether they are in fact picking up on a separate ‘London effect’.

*Hypothesis 4: Disadvantaged pupils perform better in schools with an average number of pupils in the cohort*

We find that:

- The coefficient for the KS4 cohort size is negative (-£114; std. error = £33, $p < 0.001$). As the cohort size variable was standardised, this suggests that earnings are at their peak when the cohort size is approximately average (i.e. when the standardised value is zero). Deviations from zero give positive values when squared, which translate into reductions in earnings when multiplied by the negative coefficient.
- The interaction term with squared IDACI score has a coefficient close to zero, with a *p* value very much greater than 0.05

These findings suggest that the pupils in our sample earned more if they attended a secondary school with an average-sized cohort (as opposed to a small or large cohort). We cannot conclude that this relationship is moderated by disadvantage from these results. This possibility could be explored further by including other interaction terms in the model (such as between squared cohort size and linear, non-squared IDACI score).

*Hypothesis 5: Greater secondary school spending on 'other staff' is associated with better outcomes for pupils who have ever had a special educational need or disability*

We find that:

- The longer ago a pupil was registered as having a special educational need, the higher their future earnings. An increase of one standard deviation is associated with an increase in earnings of £77 on average (std. error = 6, $p < 0.0001$)
- We cannot conclude from these results whether this relationship is moderated by spending on Other Staff, but the results suggest that any such effect is likely to be very small (coeff. = 1, std. error = 5, $p >> 0.05$).

It is hard to tell if these findings indicate that there genuinely is no relationship between Other staff spending and SEN pupil future earnings, or whether they are methodological artifacts. In particular, as the PQL model cannot handle missing variables (and as we wanted to capture the inherently ordinal nature of the Ever SEN variable), we coded pupils who had *never* been SEN as a 20 in the Ever SEN variable. In other words, pupils who have never been SEN appear in the data as if they were registered as SEN 20 years prior to their GCSE's (an arbitrarily high number). An alternative version of the model in which Ever SEN were converted to a categorical variable (with ‘never SEN’ as the reference category) might provide a more easily interpretable result.

*Hypothesis 6: Primary school spending on admin staff (including business managers and bursars) is associated with better performance for all pupils, and the relationship is moderated by disadvantage*

We find:

- At an average level of disadvantage (i.e. standardised IDACI score = 0), an increase in spending on admin staff of one standard deviation is non-significantly associated with an increase in per-pupil yearly earnings of £55 (std. error = 46, *p* = 0.23)
- Similarly, we find a positive, non-significant interaction between IDACI score and admin staff spending (coeff. = 47.92, std. error = 39, *p* = 0.22). For a given level of disadvantage (i.e. IDACI score), one standard deviation increase in E05 spending is equal to: $\beta_2 + \beta_3 x_1 = 54.95 + 47.92 * x_1$**,** where $\beta_2$is the KS2 E05 coefficient, $\beta_3$is the interaction coefficient, and $x_1$is the IDACI score. For example, when IDACI score is equal to 1 standard deviation above the mean, an increase in E05 spending of one standard deviation is associated with a £103 increase in per-pupil yearly earnings on average. When a pupil’s IDACI score is -1 (indicating that they live in an area that is more affluent than average), then a one standard deviation increase in E05 spending is associated with an increase in earnings of just £7.

The direction of these findings support our hypothesis, but the high *p* values indicate that we cannot reject the null hypothesis with these results alone. A larger sample size would be necessary in order to reject the null hypothesis.

*Hypothesis 7: There is a nonlinear relationship between primary spending on staff development & training, such that spending an average amount is associated with better outcomes than spending a large or a small amount. This relationship is moderated by disadvantage.*

We find:

- Neither KS2 E09 spending nor its interaction with IDACI score is significant (coeffs 2712 and -9697 respectively; std. errors 8381 and 6853; *p* >> 0.05)
- The direction of these coefficients suggests that increases in KS2 E09 spending in this sample are associated with *higher* earnings for pupils with average or below average IDACI scores (i.e. pupils in relatively affluent areas), and with *lower* earnings for pupils with higher IDACI scores.
- The fact that the *p* values are high means that we cannot conclude from these findings that the corresponding coefficients in the population are nonzero.

These results suggest that we cannot reject the null hypothesis. From this sample, we cannot conclude whether or not spending on staff development and training affects pupils' future earnings, and whether any such relationship is moderated by disadvantage.

*Hypothesis 8: Greater capital spending (e.g. on school improvements) is associated with better long-term earnings*

We find:

- There is a small, statistically-significant association between secondary school capital spending and future pupil earnings, such that an increase of one standard deviation in spending is associated with an average earnings increase of £244 per pupil per year (std. error = 58, *p* < 0.0001)

This result suggests that we can reject the null hypothesis and conclude that spending on capital projects is associated with higher future pupil earnings.

*Hypothesis 9: Spending more on utility bills at both primary and secondary school is associated with lower future pupil earnings. This effect is particularly strong for disadvantaged pupils.*

We find:

- There is a small, non-significant, negative effect of utilities spending on future pupil earnings, for both primary and secondary schools (coeffs = -17 [primary] and -130 [secondary]; std. errors = 44 and 82, *p* = 0.70 and 0.11). The *p* values suggest that the effect of secondary school utilities spend is perhaps more likely to be present in the population, but a larger sample would be required to confirm or reject this conjecture.

- Similarly, there are non-significant negative moderating effects of IDACI on both of these relationships (coeffs = -61 and -31, std. errors = -43 and -48, *p* = 0.15 and 0.52)
- Within *this sample*, these results indicate that, for pupils with an IDACI score one standard deviation above average, an increase in secondary school utilities spending of one standard deviation is associated with a decrease in earnings of £161 per pupil per year on average.
- For pupils with an IDACI score one standard deviation below average, the corresponding earnings decrease is £99
- Thus we find that, in this sample, higher spending on utility bills is associated with poorer outcomes, especially for disadvantaged pupils. The high *p* values indicate that we do not have sufficient evidence to conclude that these effects are also present in the population, however.

As such, we cannot reject the null hypothesis, and cannot conclude that greater spending on utilities is associated with lower earnings for pupils later in life.

*Summary*

A summary of the results of the hypothesis testing stage is presented in the table below.

| | Hypothesis | Reject the null |
|---|---|---|
| 1 | Maths KS2 performance → better earnings, especially for girls | ✓ |
| 2 | Maths KS2 performance → better earnings, especially for disadvantaged pupils | ✓ |
| 3a | Ethnic diversity → greater pupil earnings | ✓ |
| 3b | 2a moderated by ethnicity | ✗ |

| | | |
|---|---|---|
| 4a | Moderate KS4 cohort size → better earnings than small or large cohorts | ✓ |
| 4b | 3a moderated by IDACI | ✗ |
| 5 | KS4 Other staff spend → greater earnings for SEN pupils | ✗ |
| 6 | KS2 Admin staff spend → higher earnings, moderated by IDACI | ✗ |
| 7 | Staff training ^2 → lower earnings (average spend is optimal) | ✗ |
| 8 | Capital spend → higher earnings | ✓ |
| 9 | Utilities spend → lower earnings, moderated by IDACI | ✗ |

## Conclusions

In this study, we have:

- Explored the variation in school spending and pupil earnings across the country
- Investigated the extent to which the early-career earnings (or economic contribution) or individual pupils and entire cohorts can be forecasted up to 8 years in advance, using novel machine learning techniques
- Explored which pieces of information are most useful when making these predictions
- Unpicked complex, non-linear descriptive interactions between pupil and school characteristics
- Generated nine data-driven hypotheses relating to the causes of higher early career earnings for pupils from poorer backgrounds
- Tested these hypotheses using robust multilevel modelling techniques

Unsurprisingly, we have found that the picture is complex, and it would never be possible to find comprehensive answers to the social mobility puzzle with a single study. However, we have found some concrete results, and in doing so contributed to the literature on the role of education in promoting social mobility.

*Predicting pupil earnings*

The main findings from the machine learning phase of the project were (a) that it is very difficult to forecast individual pupil's earnings using administrative data alone, with a maximum R square of 0.37, and (b) that machine learning approaches seem to be significantly more effective at this task than traditional statistical models: compare the R square of 0.3 obtained by the XGboost model with the R square if 0.09 from the multilevel model.

In terms of DfE policy applications, the school-level model may be more promising than the pupil-level model. When aggregated in this way, the model's predictions of a school's average KS4 cohort earnings were accurate to within around £900 per pupil on average. Depending on cohort size, this means that we could use the model to predict the early-career economic contribution of a school's KS4 cohort to within around +/- 5%. This might enable DfE to produce indicative social mobility forecasts in a way that has never previously been possible, which in turn could promote a policymaking approach that is more proactive than reactive.

Given that most DfE activity is performed at the school, local authority or national level, predictions at an aggregated level could have powerful applications for policy. For example, we could use the model to identify schools with upcoming cohorts that are predicted to have low earnings in the future, and provide them with extra funding for careers advisors, curriculum support, one-to-one tuition and so on.

*Explaining pupil earnings*

This study also contribute to the field of research on the causes of early-career earnings. Our main finding is that there are substantial interaction effects between KS2 Maths performance and both gender and IDACI on early-career earnings. Girls and pupils from disadvantaged areas both appear to benefit significantly more from higher KS2 Maths performance than boys and pupils from affluent areas respectively.

Strikingly, the effect of IDACI on later earnings is completely eliminated for pupils with very high Maths scores (3 standard deviations above the mean).

These findings might have a number of policy implications. Firstly, they imply that initiatives to raise early Maths standards across all pupils might provide one of the most powerful ways to improve social mobility. If the act of simply doing well in KS2 Maths exams in itself has a causal impact on later earnings, then this alone could have a substantial positive effect on narrowing the earnings gap between people from poorer and wealthier backgrounds.

However, it might also be the case that it is not the exams themselves that are causing the earnings difference, but the *reaction* of pupils, parents and teachers to those exams (the pigeon-holing and stereotyping effect). More research may be necessary in order to test whether this is the case. For example, various tests of unconscious bias (e.g. Implicit Association Tests) could be carried out amongst teachers and parents to uncover their implicit attitudes towards disadvantaged and affluent children who do badly at KS2; lesson observations could be analysed qualitatively to test for potential differences in treatment of disadvantaged and affluent children with lower prior attainment, and so on.

If more support were found for this hypothesis, then a further set of policy recommendations might become relevant. Fairly conservative options could include, for example, improving unconscious bias training for teachers, or providing extra state-funded support/tuition for disadvantaged children with poor KS2 attainment. More radical options might include scrapping KS2 testing entirely (to avoid pupils being pigeon-holed).

In addition to these headline results, we also find that pupils tend to earn more on entering the labour market if they went to a secondary school with a moderately sized cohort, an ethnically diverse pupil intake, and/or a recent history of capital investment in school improvements. All of these findings would require further testing and replication before being used to directly influence policymaking. For example, experimental or quasi-experimental methods could be used to more rigorously test these effects for causality, and separate out possible confounding factors.

Finally, this study provides some evidence for the benefits to the field of education research in combining machine learning and traditional statistical methodologies. This paper has demonstrated one novel way in which the two disciplines can be combined in this context, and we suggest that an exploration of other innovative applications could provide a fruitful avenue for future research.

## References

Allen, R. (2018). *The Pupil Premium is Not Working: Do not measure attainment gaps.* Accessed 15/03/2019 at https://rebeccaallen.co.uk/2018/09/10/the-pupil-premium-is-not-working/

Altonji, J. G., E. Blom, and C. Meghir (2012). Heterogeneity in human capital investments: high school curriculum, college major, and careers. Annual Review of Economics 4 (1), 185–223

Baron, S., Field, J., & Schuller, T. (2000). *Social Capital: Critical Perspectives.* Oxford University Press, Oxford.

Richard Ernest Bellman; Rand Corporation (1957). *Dynamic programming.* Princeton University Press. p. ix. ISBN 978-0-691-07951-6.

Bergstra, J., Yamins, D., & Cox, D. D. (2013). *Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures.*

Bhutoria, A. (2016). *Economic Returns to Education in the United Kingdom.* Government Office For Science. Accessed 15/03/2019 at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/593895/Economic_Returns_To_Education_-_final.pdf

Burgess, S. (2014). Understanding the success of London's schools. Centre for Market and Public Organisation (CMPO) Working Paper, 14, 333.

Burgess, S (2016). *Human Capital: The State of the Art in the Economics of Education.* Institute of Labour Economics discussion papers, accessed 15/03/2019 at https://www.iza.org/publications/dp/9885/human-capital-and-education-the-state-of-the-art-in-the-economics-of-education

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich and Yuan Tang (2017). *xgboost: Extreme Gradient Boosting.* R package version 0.6.4.6. https://github.com/dmlc/xgboost

Codiroli Mcmaster, N. (2017). Who studies STEM subjects at A level and degree in England? An investigation into the intersections between students' family background, gender and ethnicity in determining choice. British Educational Research Journal, 43(3), 528-553.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *Equality of educational opportunity* [summary report (Vol. 2)]. US Department of Health, Education, and Welfare, Office of Education.

Crawford, C., Johnson, P., Machin, S., & Vignoles, A. (2011). *Social Mobility: A Literature Review*. Department for Business, Innovation and Skills. Accessed 15/03/2019 at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/32111/11-750-social-mobility-literature-review.pdf

Crawford, C., Gregg, P., Macmillan, L., Vignoles, A., & Wyness, G. (2016). *Higher Education, career opportunities, and intergenerational inequality*. Oxford Review of Economic Policy, vol. 32 (4). Accessed 15/03/2019 at https://academic.oup.com/oxrep/article/32/4/553/2236521

Dearden, L., Machin, S., & Reed, H. (1997). *Intergenerational mobility in Britain*. The Economic Journal, 107(440), 47-66.

DfE (2016). *Schools National Funding formula*. gov.uk. Accessed 15/03/2019 at https://consult.education.gov.uk/funding-policy-unit/schools-national-funding-formula2/supporting_documents/NFF_Stage2_schools_consultationdoc.pdf

DfE (2017). *Improving social mobility through education*. gov.uk. Accessed 15/03/2019 at https://www.gov.uk/government/publications/improving-social-mobility-through-education

DfE (2018). *Our priorities*. gov.uk. Accessed 15/03/2019 at https://www.gov.uk/government/organisations/department-for-education/about#our-priorities

Diener, E., Sandvik, E., Seidlitz, L., & Diener, M. (1993). *The relationship between income and subjective well-being: Relative or absolute?* Social Indicators Research, vol. 28 (3). https://link.springer.com/article/10.1007/BF01079018

Diener, E., & Biswas-Diener, R. (2002). *Will Money Increase Subjective Wellbeing?* Social Indicators Research, vol. 57 (2). https://link.springer.com/article/10.1023/A:1014411319119

Education Data Lab (2019). Looking at the London Effect five years on: Part one. Accessed 20/09/2019 at https://ffteducationdatalab.org.uk/2019/08/looking-at-the-london-effect-five-years-on-part-one/

Education Endowment Foundation (EEF)(2018). *The Attainment Gap*. Accessed 15/03/2019 at https://educationendowmentfoundation.org.uk/public/files/Annual_Reports/EEF_Attainment_Gap_Report_2018.pdf

Epperson, D. C. (1963). Some interpersonal and performance correlates of classroom alienation. The School Review, 71(3), 360-376.

Farkas, G. (1996). *Human Capital or Cultural Capital? Ethnicity and Poverty Groups in an Urban School District*. Routledge, New York: 2017. DOI: https://doi.org/10.4324/9780203789575

Friedman, J. H. (2001). *Greedy function approximation: a gradient boosting machine*. Annals of statistics, 1189-1232.

Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. Statistics in medicine, 22(9), 1365-1381.

Hanushek, E. A. (2008). *Education production functions*. The new Palgrave dictionary of economics. Basingstoke: Palgrave Macmillan

Hayward H., Hunt E., Lord A. (2014), '*The economic value of key intermediate qualifications: estimating the returns and lifetime productivity gains to GCSEs, A levels and apprenticeships*', Department for Education

Israni, E. T. (2017). *When an Algorithm Helps Send You to Prison.* New York Times. Accessed 15/03/2019 at http://www.cchsenglish.com/uploads/3/1/5/8/3158478/when_an_algorithm_helps_send_you_to_prison_-_the_new_york_times.pdf

Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2018). *caret: Classification and Regression Training.* R package version 6.0-80. https://CRAN.R-project.org/package=caret

McLaughlin, J. E., McLaughlin, G. W., McLaughlin, J. S., & White, C. Y. (2016). Using Simpson's diversity index to examine multidimensional models of diversity in health professions education. *International journal of medical education*, 7, 1–5. doi:10.5116/ijme.565e.1112

Lessof, C., Ross, A., Brind, R., Harding, C., Bell, E., & Kyriakopoulos, G. (2018). *Understanding KS4 attainment and progress: Evidence from LSYPE2.* Department for Education. Accessed 15/03/2019 at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/748514/Understanding_KS4_LSYPE2_research-report.pdf

Levin, J. (2002). For whom the reductions count: A quantile regression analysis of class size and peer effects on scholastic achievement. In Economic Applications of Quantile Regression (pp. 221-246). Physica, Heidelberg.

Masci, C., Johnes, G., & Agasisti, T. (2018). *Student and school performance across countries: A machine learning approach.* European Journal of Operational Research, vol. 269.

National Audit Office (2018). *Converting Maintained Schools to Academies.* Accessed 20th September 2018 at https://www.nao.org.uk/report/converting-maintained-schools-to-academies/

George Psacharopoulos & Harry Anthony Patrinos (2018) *Returns to investment in education: a decennial review of the global literature*, Education Economics, 26:5, 445-458, DOI: 10.1080/09645292.2018.1484426

R Core Team (2017). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schiltz, F., Masci,C., Tommaso Agasisti & Daniel Horn (2018) Using regression tree ensembles to model interaction effects: a graphical approach, Applied Economics, 50:58, 6341-6354, DOI: 10.1080/00036846.2018.1489520

Rist, R. (1970). Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. Harvard educational review, 40(3), 411-451.

Sharp, C., Macleod, S., Bernadelli, D., Skipp, A., Higgins, S. (2015). *Supporting the attainment of disadvantaged pupils.* National Foundation for Educational Research; Ask Research; Durham University. Accessed 15/03/2019 at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/473976/DFE-RS411_Supporting_the_attainment_of_disadvantaged_pupils_-_briefing_for_school_leaders.pdf

Singh, A., Thakur, N., Sharma, A. (2016). *A Review of supervised machine learning algorithms.* 3rd International Conference on Computing for Sustainable Global Development, https://ieeexplore.ieee.org/abstract/document/7724478

Social Mobility Commission (2019). State of the Nation, 2018-19: Social Mobility in Great Britaion. Accessed 20/09/2019 at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/798404/SMC_State_of_the_Nation_Report_2018-19.pdf

Uğ uz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems, 24(7), 1024-1032.

Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S.* Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'.* R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Willard, J., & Madon, S. (2016). Understanding the connections between self-fulfilling prophecies and social problems. In Interpersonal and Intrapersonal Expectancies (pp. 117-124). Routledge.

Yachen Yan (2016). *rBayesianOptimization: Bayesian Optimization of Hyperparameters.* R package version 1.1.0. https://CRAN.R-project.org/package=rBayesianOptimization