Title

Predicting Ofsted Inspection Grades Using School Performance, Demographic, and Prior Inspection Data

Executive Summary

This project investigated whether publicly available data could be used to predict Ofsted inspection grades for secondary schools in England. The aim was to evaluate the extent to which factors such as academic performance, pupil demographics, and previous inspection outcomes correlate with and help predict Ofsted ratings.

Three datasets were merged: the most recent Ofsted inspection outcomes, school-level performance tables (including academic attainment and pupil progress), and school characteristics (including demographics and special educational needs data).

Several supervised learning models were trained and evaluated using accuracy, precision, recall, and weighted F1-score. The most accurate model achieved 74% accuracy and a weighted F1-score of 73% (that balances precision and recall). The models showed the strongest performance in predicting schools rated 'Good'.

However, the models failed to reliably predict schools rated 'Inadequate', which limits their usefulness for policy or early-warning interventions and therefore is not suitable for practical deployment in its current form.

The most predictive features included previous inspection subgrades, particularly leadership and overall effectiveness, average GCSE performance across eight subjects including English and Maths, average prior attainment, and the proportion of pupils with special educational needs. Conversely, lower Ofsted grades were most strongly associated with, lower pupil progress scores (especially in maths and English Baccalaureate areas), and greater underperformance relative to national benchmarks.

In conclusion, while the modelling approach showed potential, its inability to flag the most vulnerable schools makes it unsuitable for real-world deployment without significant refinement. Future iterations could focus on addressing class imbalance and incorporating additional qualitative or contextual data.

Introduction

In England, Ofsted ratings have far-reaching consequences for schools, affecting parental choice, funding, and staff morale. However, inspection frequency varies and the process can be resource intensive. I sought to answer the question: Can school inspection outcomes be reliably predicted using publicly available data? This question is well-suited for a data-driven approach given the standardised nature of school performance and characteristics data collected annually by the Department for Education. Predictive modelling may offer a scalable means of flagging schools at risk of poor ratings or highlighting those due for reinspection.

While Ofsted inspections also consider lesson observations, pupil interviews, safeguarding practices, school leadership, and broader contextual factors, this project focuses on structured data alone. The aim is not to take into account the full inspection elements but to consider how far a standardised dataset can still provide meaningful predictive insight.

Methods

Methods: Data Sources

Three primary data sources were used:

1. **Ofsted Inspection Outcomes (March 2025)** – Published by Ofsted and accessed through the UK Government data service, this dataset includes inspection grades and subdomain scores (e.g. leadership, safeguarding).

   https://www.gov.uk/csv-preview/680117c990dd6a0497e28815/Management_information_-_state-funded_schools_-_latest_inspections_-_as_at_31_Mar_2025.csv

2. **School Performance Tables** – From the Department for Education's national statistics, these include academic performance measures such as Attainment 8, Progress 8, EBacc entry proportions, and prior attainment.

3. **School Characteristics Dataset** – This contains pupil demographic information including proportions of students eligible for free school meals, English as an Additional Language (EAL), special educational needs (SEN), and contextual deprivation (IDACI).

   https://explore-education-statistics.service.gov.uk/data-catalogue?publicationId=c8756008-ed50-4632-9b96-01b5ca002a43&releaseVersionId=b76a938a-7875-4542-af20-0b23ecb99a49&themeId=74648781-85a9-4233-8be3-fe6f137165f4

Data Cleaning and Feature Engineering:

1. Converted object-type columns to numerical formats where applicable.
2. Checked for and removed duplicate records.
3. Calculated the number and percentage of missing values per feature to guide imputation.
4. Retained only rows summarised at the whole-school level.
5. Removed rows missing URNs or target labels.
6. Merged datasets using the unique school identifiers (URNs).
7. Cleaned invalid or placeholder codes (e.g. '9', '9.0'); standardised data types.
8. Using domain knowledge, selected relevant features from academic results, pupil demographics, inspection history, and contextual variables.
9. Included prior subgrades from previous inspections (e.g. leadership, personal development) as predictors.
10. Converted categorical variables using appropriate encoding methods (e.g. one-hot encoding).
11. Checked feature distributions and addressed right-skewed variables.
12. Outliers were assessed visually and considered in modelling decisions.
13. Calculated feature–target correlations to prioritise predictive features.
14. Removed highly collinear features ($r > 0.85$) to reduce redundancy in regression models.

15. Used median imputation for missing numeric features (Simple Imputer).
16. Balanced class distributions to address imbalanced outcome labels.
17. Split the data into training, validation, and test sets using stratified sampling.
18. The project was split into two separate notebooks: one for logistic regression and one for tree-based models, since the preprocessing requirements differed significantly. For example, feature correlation checks were necessary for the regression model but not for tree-based models. Additionally, tree-based models did not require feature scaling or transformation of skewed variables and were ok with outliers and multicollinearity.

Model Selection and Justification

I experimented with several supervised classification models including logistic regression, random forest, XGBoost, and LightGBM. Tree-based ensemble models were chosen for their ability to handle feature interactions and non-linearity. Logistic regression served as a baseline due to its interpretability.

Model performance was evaluated using accuracy, precision, recall, and weighted F1-score. The best-performing model (LightGBM) achieved 74% accuracy and a weighted F1-score of 73%. However, it failed to identify schools with the lowest Ofsted grade, which reduces its practical application.

Hyperparameter tuning was carried out for LightGBM and XGBoost using GridSearchCV which systematically explores combinations of hyperparameter values through cross-validation to optimise model performance. Random Forest was retained as a baseline and not tuned, as performance lagged behind the boosted models.

To address the issue of class imbalance, particularly the underrepresentation of 'Inadequate' ratings, I applied SMOTE **(**Synthetic Minority Oversampling Technique) to artificially increase the number of samples in minority classes during training to help the model generalise better to rare categories.

These methods reflect a standard machine learning pipeline suitable for multiclass classification of ordinal labels and were designed to maximise predictive performance while maintaining interpretability.

Results

Model performance varied across classification types, with tree-based models (especially LightGBM) outperforming logistic regression in terms of accuracy and class-level recall. The best-performing model achieved 74% accuracy and a weighted F1-score of 73%. However, it consistently struggled to identify schools rated 'Inadequate', a critical shortfall in terms of practical policy application. Visualisations in the accompanying Jupyter notebooks illustrate model performance, class distributions, and confusion matrices. These confirm that the models predominantly predicted the majority class ('Good'), with decreasing performance for 'Outstanding' and 'Requires Improvement', and little success in flagging 'Inadequate' schools.

Feature importance analyses and correlation visualisations revealed that the strongest positive predictors of Ofsted grade were previous inspection scores for leadership and overall effectiveness, average Key Stage 2 attainment, GCSE Attainment 8 scores (average performance across eight qualifications), and the proportion of pupils with special educational needs (SEN). Negatively correlated features included low Progress 8 scores

(indicating underperformance relative to prior attainment), poor maths progress, high persistent absence, and high unauthorised absence.

Full visual outputs, including model performance matrices and charts, feature distributions, correlation heatmaps, and importance rankings, are available in the notebooks hosted in this repository.

Conclusion

This project shows that machine learning can provide useful but limited predictions of Ofsted inspection outcomes using standardised school-level data. However, all models showed poor recall and precision for lower grades such as 'Inadequate', raising concerns about fairness and practical reliability. Despite tuning and oversampling, the modelling failed to offer meaningful predictive value for the most critical inspection outcomes. I therefore do not recommend this modelling framework for deployment in educational oversight or school accountability. It may serve a role in research or exploratory analysis, but not for identifying schools in need of urgent attention

Recommendations for next steps:

- Address class imbalance more robustly.

  Future versions of the model could explore advanced resampling techniques or use ordinal classification methods that explicitly take into account the natural order of Ofsted grades (from 'Outstanding' to 'Inadequate'). In this project, the grades were encoded as numbers, but treated as separate categories in a standard multiclass setup, without considering their rank. Ordinal classifiers could improve performance by penalising larger misclassification gaps more heavily, for example, mistaking 'Outstanding' for 'Inadequate' would be seen as more serious than predicting 'Good' instead of 'Outstanding'.

- Integrate qualitative or contextual data**.**

  Adding information such as school leadership changes, safeguarding incidents, or findings from staff and parent surveys could help the model capture elements of inspection that go beyond academic progress. Other contextual signals like pupil and staff turnover, financial health, or the type of school governance might also reveal patterns not visible in academic data alone. To go further, natural language processing techniques could be applied to past Ofsted reports to extract tone, sentiment, or commonly mentioned concerns, offering a richer view of what influences inspection outcomes.