Cabinet Working Group
Houses of Multiple Occupation/Selective Licencing

DATE	**4th March 2026**
REPORT OF	**Joanne Robinson, Assistant Director Policy Strategy and Resources**
SUBJECT	**HMO Data Science Activity**
STATUS	**Open**

CONTRIBUTION TO OUR AIMS

This report supports the Council's strategic aims by strengthening intelligence led enforcement and improving standards in the private rented sector. The data driven approach enables a more proportionate and targeted use of resources, focussing activity on properties most likely to pose risks to and poor outcomes for residents. It helps protect people living in shared accommodation, supports safer, stronger and healthier homes and communities, and provides a stronger evidence base to inform future policy decisions, including any consideration of selective licensing or other regulatory interventions.

EXECUTIVE SUMMARY

This report sets out a pilot data led approach to support the identification of small, non-licensable HMOs in North East Lincolnshire. It provides a risk ranked evidence base to inform targeted validation activity and strategic discussion, supporting more proportionate use of housing team resource. The work is intended to support professional judgement and statutory processes, and to inform future policy considerations, including any potential selective licensing.

RECOMMENDATIONS

It is recommended that the Cabinet Working group agrees to:

- Phase 1: Validate Very High and High properties (c.140) via desktop checks and targeted visits. (Resource required).

- Phase 2: Sample Medium properties to refine model precision and understand false positives.

- Intelligence Loop: Feed inspection results back into the model for calibration.

- Data Enhancement: Incorporate Experian household composition data and cross-council intelligence (Fraud Team, LoCTA system) as available.

- Governance: Improve UPRN coverage and address data hygiene to support accuracy and future analysis.

REASONS FOR DECISION

The recommendations support a proportionate and evidence-led approach to targeting potential non-licensable HMOs, enabling resources to be focused on properties most likely to present risk.

A phased validation process mitigates the risk of misinterpretation, strengthens data quality, and ensures that professional judgement and statutory processes remain central. The approach improves intelligence, supports better governance, and provides a stronger evidence base to inform any future policy considerations, without committing the Council to further selective licensing at this stage.

1. BACKGROUND AND ISSUES

Purpose and Approach

NELC has piloted a data-driven method to identify small Houses in Multiple Occupation (HMOs) with four or fewer occupants, which currently fall outside mandatory licensing. The aim is to create a risk-ranked list of properties for targeted validation and enforcement, and to strengthen the evidence base for future policy decisions.

A machine-learning model (Random Forest) was developed using council, national, and third-party datasets. Each property is assigned a probability of being an HMO-like dwelling, resulting in a prioritised list grouped into likelihood bands: Very High, High, Medium, Low, and Very Low.

The work is intended to support strategic and operational decision-making and does not replace statutory processes, professional judgement or service-led validation.

Data Sources and Engineering

Local Authority Data

- Council Tax accounts (ownership indicators)
- Electoral Roll (multi-household occupancy signals)
- NELC HMO Licensing Register (used for model training)
- Planning & Building Control (conversion/subdivision signals)
- Supported Accommodation,
- LLPG
- Environmental complaints

National & Third-Party Data

- ONS Census 2021 (contextual, not property-level)
- Public energy efficiency data (property characteristics)

Addresses were standardised and linked via UPRN to unify data. Features engineered included room counts, floor area, ownership types, and

neighbourhood indicators. Data quality constraints were noted, especially in distinguishing unrelated adults from family households.

Modelling and Insights

- The Random Forest model was trained on engineered features, balancing classes and constraining depth for robustness.
- Evaluation used ROC-AUC and precision-recall, with thresholds set for likelihood bands.
- Crime rate was found to be a weak predictor locally and was removed from later model iterations.
- Room counts and heated rooms were stronger predictors; property type was less useful due to data skew.

Outputs

- Properties are ranked and grouped: Very High (60), High (81), Medium (94), Low (93), Very Low (938).

- These bands represent a small proportion of the borough's housing stock and privately rented households.

- A Power BI map and summary view support operational targeting, with ward-level hotspot mapping for authorised officers.

- Ward-level outputs are provided for contextual understanding and operational prioritisation only and do not constitute place-based designation or policy proposals.

Implications for future selective licensing consideration

Subject to the outcome of phased validation activity, the emerging intelligence suggests that any future consideration of selective licensing should remain tightly targeted and proportionate, building on the rationale previously agreed by Cabinet for parts of East Marsh.

The original designation was based on a convergence of factors, including very high levels of deprivation, significant concentrations of private rented accommodation, persistent poor housing conditions, and associated impacts on community wellbeing, where other interventions had not delivered sufficient improvement.

The current data-led analysis does not in itself propose any new designated areas; however, it indicates that a high proportion of properties assessed as having a strong likelihood of being small, non-licensable HMOs are concentrated within a small number of inner-urban wards adjacent to East Marsh, notably Sidney Sussex, Heneage and Park.

If validation confirms these patterns, this would suggest that any future selective licensing proposal should focus on clearly defined streets or micro-areas where

high private rented sector density, concentrations of higher-risk shared accommodation and wider neighbourhood impacts coincide, rather than adopting a borough-wide approach, and would be subject to further evidence gathering, consultation and separate Cabinet consideration.

2. RISKS, OPPORTUNITIES AND EQUALITY ISSUES

There is a risk that data-led analysis could be misinterpreted if used in isolation; this is mitigated through phased validation, professional judgement and statutory processes. The approach creates opportunities to improve targeting, reduce reactive enforcement and make more proportionate use of resources. By focusing on properties most likely to present risk, the work has the potential to improve housing conditions and support vulnerable residents and communities who may be disproportionately affected by poor quality or overcrowded accommodation.

3. OTHER OPTIONS CONSIDERED

Discussions regarding existing reactive and complaint led enforcement were considered but the recommendation for a data led approach was preferred.

4. REPUTATION AND COMMUNICATIONS CONSIDERATIONS

There are potential reputational risks if data led analysis is misunderstood as enforcement action or a commitment to licensing. This will be mitigated through clear internal governance, phased validation and appropriate communications that emphasise the pilot nature of the work and the continued role of professional judgement and statutory processes. The approach also presents positive reputational opportunities by demonstrating that the Council is taking a proportionate, evidence led and responsible approach to improving housing standards for residents and communities.

5. FINANCIAL CONSIDERATIONS

The data science activity set out in this report has been and will continue to be delivered within existing resources and does not commit the Council to any additional expenditure at this stage. The recommendations for targeted checks via the housing team may mean a reprioritisation of workload or additional resource requirements.

Any future policy, project or licensing proposals arising from this work would be subject to separate financial appraisal and decision making.

6. CHILDREN AND YOUNG PEOPLE IMPLICATIONS

There are no direct impacts on children and young people. Improved targeting of poor housing conditions may have indirect benefits for children and young people living in the private rented sector.

7. CLIMATE CHANGE, NATURE RECOVERY AND ENVIRONMENTAL IMPLICATIONS

There are no direct climate change, nature recovery or environmental impacts arising from this report.

8. PUBLIC HEALTH, HEALTH INEQUALITIES AND MARMOT IMPLICATIONS

Poor quality and overcrowded housing is associated with adverse physical and mental health outcomes and can disproportionately affect lower-income households and vulnerable residents. By supporting more targeted and proportionate action to identify and address higher-risk accommodation, the approach has the potential to contribute to improved living conditions and reduced health inequalities. The work aligns with Marmot principles by focusing on prevention and the wider social determinants of health.

9. FINANCIAL IMPLICATIONS

This is a technical report. There are currently no financial implications based on the detail in this report. Any future policy or service changes arising from this work would be subject to a separate financial appraisal and decision.

10. LEGAL IMPLICATIONS

This is a technical report. There are currently no financial implications based on the detail in this report. Any future policy or service changes arising from this work would be subject to separate legal appraisal and decision making.

11. HUMAN RESOURCES IMPLICATIONS

This is a technical report. There are currently no financial implications based on the detail in this report. Any future policy or service changes arising from this work would be subject to a separate human resource appraisal and decision.

12. WARD IMPLICATIONS

The analysis highlights variation in the distribution of potential non-licensable HMOs across wards, which is intended to inform understanding and operational prioritisation only. There are no direct ward specific implications or designations arising from this report, and any future place-based proposals would be subject to separate consideration.

13. BACKGROUND PAPERS

Report: IDENTIFYING NON-LICENSABLE (≤4 PERSON) HMOS IN NORTH EAST LINCOLNSHIRE USING DATA SCIENCE (inc visual map)

14. CONTACT OFFICER(S)

Drew Hughes. Head of Strategy, Policy and Performance
Paul Silvester. Insights - Data Science

Joanne Robinson
Assistant Director of Policy Strategy and Resources

IDENTIFYING NON-LICENSABLE (≤4 PERSON) HMOS IN NORTH EAST LINCOLNSHIRE USING DATA SCIENCE

Prepared by: Insights, Data Science.

Meeting: Cabinet Working Group/Select Committee

Date: 04 March 2026

1. EXECUTIVE SUMMARY

North East Lincolnshire Council (NELC) is trialing a data-driven approach to identify small Houses in Multiple Occupation (HMOs) that currently fall outside mandatory licensing (≤4 occupants). The goal is to produce an intelligence-led, risk-ranked list of properties for proportionate validation and enforcement, while strengthening the evidence base for potential policy options.

Using council, national and third-party datasets, the Insights team has built a machine-learning model (Random Forest) that assigns each candidate property a probability of being an HMO-like dwelling.

Initial results provide a prioritised list of properties grouped into Very High, High, Medium, Low and Very Low likelihood bands. Current distribution: Very High (61), High (87), Medium (77), Low (61), Very Low (948).

This report summarises context, data inputs, modelling approach, outputs, limitations, and next steps.

2. BACKGROUND AND POLICY CONTEXT

Smaller HMOs (≤4 occupants) can impact local amenity, standards, and neighbourhood stability but are not currently subject to mandatory licensing. Understanding their distribution helps target advice, engagement, and enforcement.

Evidence papers from other authorities (e.g., Sutton, Ipswich, Hyndburn) show useful techniques for estimating “hidden HMOs” and illustrate how intelligence can support Article 4 / selective licensing decisions. We have reviewed these to benchmark methodology and presentation standards for our local context.

3. DATA SOURCES

We integrated multiple datasets to create a unified, property-level view (standardised via address matching and UPRN where possible).

3.1 Local authority datasets

- Council Tax accounts. Indicators such as non-resident liable parties, company/agent ownership.
- Electoral Roll (geocoded). Aggregated electors per address to suggest possible multi-household occupancy (acknowledging data quality caveats).
- NELC HMO Licensing Register (5+). Used as the "positive class" for modelling; includes persons on license where present.
- Planning & Building Control. Signals of subdivision/conversion (C4 use, HMOs, flats over shops).
- Supported Accommodation addresses; LLPG; Environmental complaints (noise, waste, parking). Contextual indicators for multi-occupation and neighbourhood effects.

A working log of the HMO project's data sources, owners and access dates is maintained (e.g., CT liable party lists, geocoded electoral register, complaints, LLPG, EPC data).

3.2 National & third-party datasets

- ONS Census 2021 (RM192/RM193). Household composition and HMO-related classifications used for ward- and LA-level context and benchmarking—not for property-level identification.
- Public energy efficiency data — property characteristics (floor area, rooms) as model features where available.

4. DATA ENGINEERING & INTEGRATION

4.1 Preparation and linkage

Address standardisation & UPRN. We harmonised addresses across sources and linked to UPRN where possible to reduce duplicates and enable feature joins.

Feature derivation. We engineered features including number of habitable rooms, heated rooms, total floor area, ownership/liable-party types, supported accommodation flags, planning signals, and (where appropriate) neighbourhood indicators (deprivation, crime).

Data quality constraints. For the electoral roll, we aggregated counts to address level; however, the absence of surname-level information limits the ability to distinguish family households from unrelated adults.

4.2 Training table

Given the absence of a definitive list of known non-HMOs, we created a labelled dataset by combining licensed HMOs (positives) with a "noisy negative" sample from the wider housing stock. This approach, documented in our model summary, allows the algorithm to learn "HMO-likeness" for ranking purposes, rather than definitive classification.

5. DATA SCIENCE METHODS

5.1 Model selection and training

We used a type of machine-learning model called a Random Forest, which is made up of 500 simple decision trees. Before training it, we cleaned the data, filled in missing information, converted categories into numbers, and scaled everything to a similar range. This method is good at spotting complex patterns and works well even when the data is unbalanced or a bit messy — which makes it a strong choice for ranking properties by risk.

5.2 Evaluation and thresholding

We checked how good the model is at ranking risks by using two measures: one that looks at how well it separates higher-risk from lower-risk cases overall, and another that focuses on how accurately it identifies the rare high-risk cases, then created operational thresholds to form Very High / High / Medium / Low / Very Low categories. Thresholds are adjustable and will be tuned with Enforcement as we learn from field validation.

5.3 Feature insights and iterations

Early analysis suggested crime rate was not a strong predictor for our local suspected HMO cohort and was de-emphasised/removed in later runs.

Rooms and heated rooms emerged as stronger differentiators; property type offered limited marginal value due to skew.

Where possible, we plan to enhance owner/liable-party features (e.g., distinguishing companies/agents vs. estates) once comparable data is available for both known HMOs and candidates.

6. OUTPUTS

6.1 Ranked list and bands

The current model output ranks candidate properties and groups them by likelihood. As of the latest summary, bands and counts are: Very High (61), High (87), Medium (77), Low (61), Very Low (948). We will supply the full list to Housing Enforcement for controlled operational use. (Property-level lists are not included in public documentation.)

6.2 Scale and context

In proportional terms (indicative only), these bands correspond to 0.08%–1.25% of the approximate 76,000 properties in the borough, and 0.43%–6.68% of the 14,185 privately rented households (Census 2021). This provides a screening view, not a definitive count.

The table below shows a breakdown of the list of properties by Ward and HMO Probability. 83% of all properties indicated as ‘Very High’ probability can be found within the Sidney Sussex, Heneage and Park Wards.

Ward Name\HMO Probablilty	Very High	High	Medium	Low	Very Low	Grand Total
Sidney Sussex	20	26	10	7	89	152
Heneage	19	16	28	5	83	151
Park	12	9	11	15	47	94
East Marsh	5	16	18	13	105	157
Croft Baker	3	8	6	6	91	114
West Marsh	2	12	4	7	111	136
Immingham				3	58	61
Yarborough				2	69	71
Scartho				2	41	43
Wolds				1	27	28
South					74	74
Humberston and New Waltham					50	50
Haverstoe					40	40
Waltham					36	36
Freshney					25	25
Grand Total	**61**	**87**	**77**	**61**	**946**	**1232**

6.3 Benchmarking NELC

The table below indicates potential benchmarking of small HMO activity for NELC against some local comparators. This benchmarking information is taken from the 2021 Census, which asked the question around HMOs for the first time. The Census definition is as follows;
A dwelling where unrelated tenants rent their home from a private landlord is a HMO, if both of the following apply:

- at least three unrelated individuals live there, forming more than one household
- toilet, bathroom or kitchen facilities are shared with other tenants

A small HMO is shared by 3 or 4 unrelated tenants. A large HMO is shared by 5 or more unrelated tenants

	Households of multiple occupancy (HMO)		
	Is a large HMO	Is a small HMO	Grand Total
Lincoln	0.84%	1.89%	2.73%
Kingston upon Hull	0.24%	0.51%	0.75%
Ipswich	0.08%	0.23%	0.31%
South Holland	0.03%	0.18%	0.21%
North Lincolnshire	0.02%	0.10%	0.12%
North East Lincolnshire	0.23%	0.09%	0.31%
South Kesteven	0.00%	0.07%	0.07%
North Kesteven	0.01%	0.05%	0.06%
Rutland	0.00%	0.04%	0.04%
West Lindsey	0.00%	0.04%	0.04%
East Riding of Yorkshire	0.01%	0.04%	0.05%
East Lindsey	0.01%	0.02%	0.03%

Due to the way the census is conducted and relies on answers from the respondent these figures should be treated with caution and not as an accurate proxy for HMOs, but they do offer a starting point for comparison.

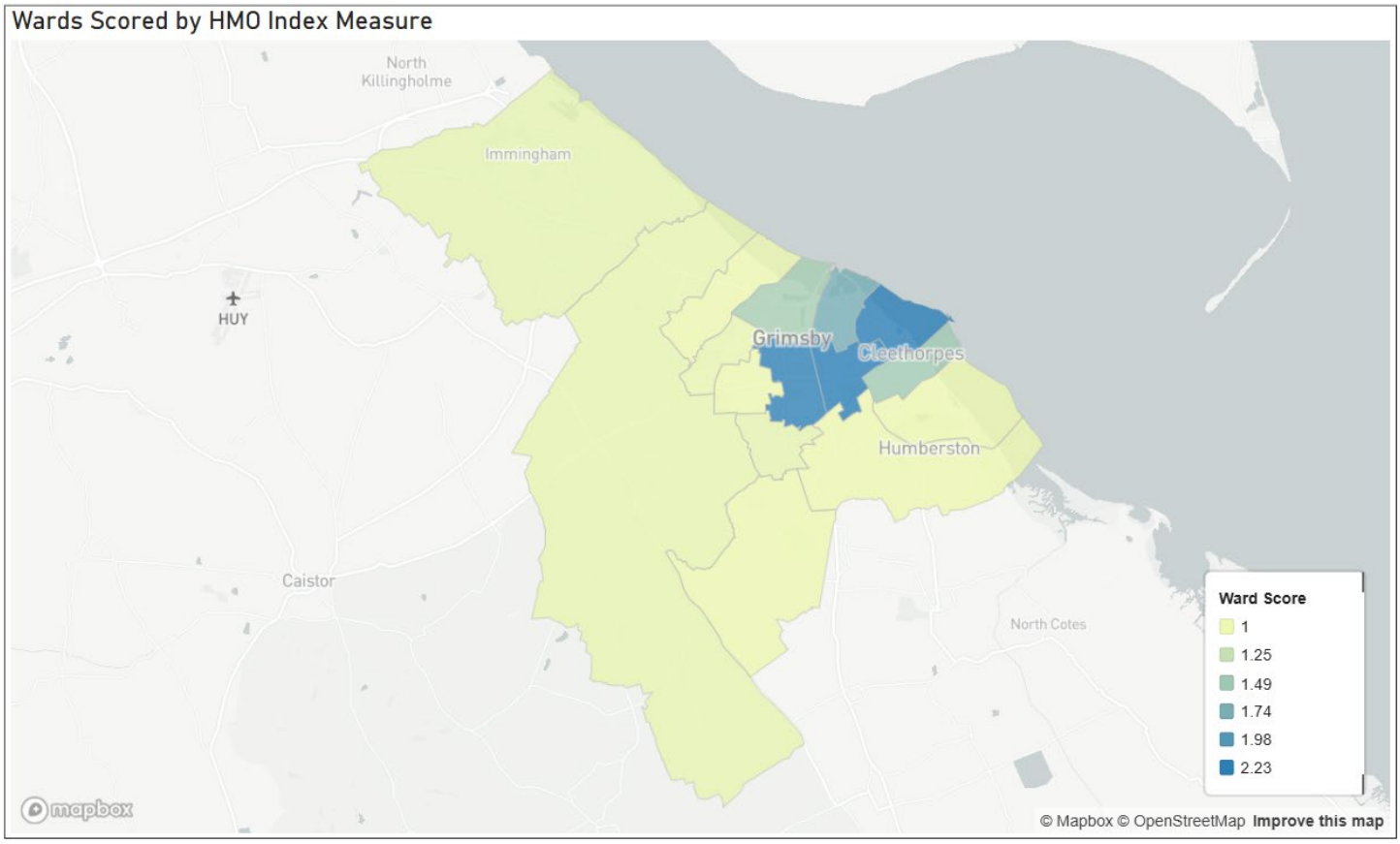
We can attempt to incorporate the data science work done into possible benchmarking. Although this is only an estimate of successful probabilities (20% success rate based on the range of 0.08%-1.25%) and only uses Census data for other authorities – this would increase the possible number of small HMO's in North East Lincolnshire to 0.23%, as shown below.

	Households of multiple occupancy (HMO)		
	Is a large HMO	Is a small HMO	Grand Total
Lincoln	0.84%	1.89%	2.73%
Kingston upon Hull	0.24%	0.51%	0.75%
North East Lincolnshire	0.23%	0.23%	0.46%
Ipswich	0.08%	0.23%	0.31%
South Holland	0.03%	0.18%	0.21%
North Lincolnshire	0.02%	0.10%	0.12%
South Kesteven	0.00%	0.07%	0.07%
North Kesteven	0.01%	0.05%	0.06%
Rutland	0.00%	0.04%	0.04%
West Lindsey	0.00%	0.04%	0.04%
East Riding of Yorkshire	0.01%	0.04%	0.05%
East Lindsey	0.01%	0.02%	0.03%

It should be noted that this benchmarking figure for NEL could be higher or lower depending on successful validation of the indicated properties and that the figures for other authorities could also change if a similar piece of work identifying HMOs was done for each locality.

6.4 Visualisations

We will maintain a Power BI map and summary view to support operational targeting and oversight, with ward-level hotspot mapping and drill-down for authorised officers. A simplified map has been created for this report.



By allocating each rank a score (Very High = 5, Very Low = 1) we can calculate a total score for each ward, which when divided by the number of properties gives a Ward HMO Index Score. This is visualized above and shown in the table below.

Ward Name	Ward Score
Heneage	2.23
Sidney Sussex	2.22
Park	2.19
East Marsh	1.75
Croft Baker	1.47
West Marsh	1.43
Immingham	1.05
Scartho	1.05
Wolds	1.04
Yarborough	1.03
South	1.00
Humberston and New Waltham	1.00
Haverstoe	1.00
Waltham	1.00
Freshney	1.00

7. RECOMMENDATIONS

7.1 Operational (Enforcement)

Phase 1 validation (Very High + High).

If further data certainty is desired, consider undertaking structured desktop checks and targeted visits for c.140 highest-ranked properties. This has a resource implication that needs to be explored further. Document outcomes to feed model calibration and threshold review.

Phase 2 sampling (Medium).

Sample-check a subset to refine precision/recall trade-offs and understand false-positive patterns.

Intelligence loop.

Capture inspection results in a standard template; feed back into model features and thresholds.

7.2 Data enhancement

Experian household composition (via Civica OnDemand). Incorporate once available to improve household structure signals; note that this data will arrive after the immediate February reporting deadline.

Cross-council intelligence working with the Fraud Team (using an external system LoCTA) to validate liable-party anomalies, subject to legal and procurement steps.

Address governance. Continue improving UPRN coverage and consistent address hygiene across systems.

7.3 Implications for future selective licensing consideration

This analysis does not propose selective licensing and does not designate any areas.

The outputs strengthen the evidence base for future strategic consideration, alongside professional judgement and statutory processes.

Higher-likelihood small, non-licensable HMOs are concentrated within a small number of inner-urban wards, particularly Sidney Sussex, Heneage and Park, rather than being evenly distributed across the borough.

These areas are adjacent to East Marsh, where selective licensing was previously agreed by Cabinet based on high private rented sector density, poor housing conditions and wider neighbourhood impacts.

Subject to validation, any future consideration of selective licensing would most appropriately focus on clearly defined streets or micro-areas where multiple risk factors align, rather than adopting a borough-wide approach, and would be subject to separate Cabinet decision-making.

8. LIMITATIONS

- The ranked list is an intelligence product that prioritises likely HMOs; it is not a register. The data shows the probability of the whereabouts of HMOs, further work will need to be undertaken to confirm the assumptions made from the data.

- Data quality and coverage. Some features (e.g., electors per address) lack granularity (no surname differentiation) and can under- or over-signal multi-household living.
- Temporal misalignment. Datasets are extracted on different dates; occupancy can change quickly.
- Ethics & privacy. Approach remains property-level; personal data is minimised per the DPIA.

9. APPENDICES

A. Dataset inventory.

B. Model summary document.

Stage	Main Technique(s)	What this part really does and why
Data combination	Positive + noisy negative concatenation	The script puts confirmed HMOs (label=1) together with suspected unlicensed HMO properties (label=0) to make one training table. This is the only practical way to create a labelled dataset when you have few confirmed positives and no perfect list of confirmed non-HMOs.
Preprocessing	Column Transformer + imputation + OHE (capped at 30) + scaling	The script cleans and standardises the data before training. It fills missing values, encodes categorical data into numbers the model can use, and scales all number columns so none dominate unfairly. Capping rare categories at 30 keeps the data from exploding in size. This step makes sure the model gets clean, consistent input every time.
Model	Random Forest (500 trees, balanced weights, max depth 12, min leaf 5)	The script uses the Random Forest algorithm to learn patterns from the data. It builds 500 decision trees, balances them so rare HMOs get proper attention, and limits how deep they grow to avoid memorising noise. The result is a model that gives reliable probability scores showing how much each property looks like a known HMO. This is a safe, strong choice for ranking tasks with messy real-world data.
Evaluation	ROC AUC + Precision-Recall curve + best F1 threshold + threshold table	The script checks how well the model ranks properties, not just how many it gets exactly right. AUC measures overall ranking quality. Precision-Recall focuses on the rare HMOs. The threshold table shows what happens at different cut-off points so the council can decide: catch more possible HMOs (more inspections) or be stricter (fewer wasted visits). These metrics match the real goal: a useful priority list.
Diagnostics	Known-set probability stats + flip heuristic	The script runs the model on the known HMOs again to test it. Their scores should mostly be high. This check makes sure the scores point in the right direction before ranking suspected properties.

Ranking	Probability sort + likelihood bins (Very Low → Very High) + rank number	The script sorts the suspected properties from highest to lowest probability. It adds a rank number (1 = most suspicious) and groups the probabilities into simple labels (Very High, High, etc.). This creates a clear, actionable list the enforcement team can use straight away to decide which properties to check first.
Interpretability	Gini feature importance + probability histograms (known vs suspected)	The script shows which features (room count, area, location, etc.) influence the decisions most. It also plots two histograms: one for known HMOs (should peak high) and one for suspected properties (should mostly be low with some higher scores). These visuals help confirm the model learned sensible patterns and make the results easier to trust and explain.

HMO Data Set	Description of the Data Set
1	List of properties and the Council Tax Liable Parties
2	List of current Supported Accomodation addresses
3	Public sumitted complaints data (eg noise, parking, litter)
4	North East Lincolnshire Postcodes list
5	Geocoded Electoral Register
6	List of Current HMO Properties
7	Building Control & Planning Data
8	Local Land and Property Gazetteer
9	Property Energy Efficiency Data
10	
11	
12	
13	
14	
15	

Service	Data extracted	Any restrictions on use
CT&B	13/10/2025	
CT&B	09/01/2026	
IT	28/01/2026	
Insights	21/01/2026	
Elections	03/02/2026	
	09/02/2026	
IT/Regen	21/01/2026	
	21/01/2026	

Notes	Date Access Provided