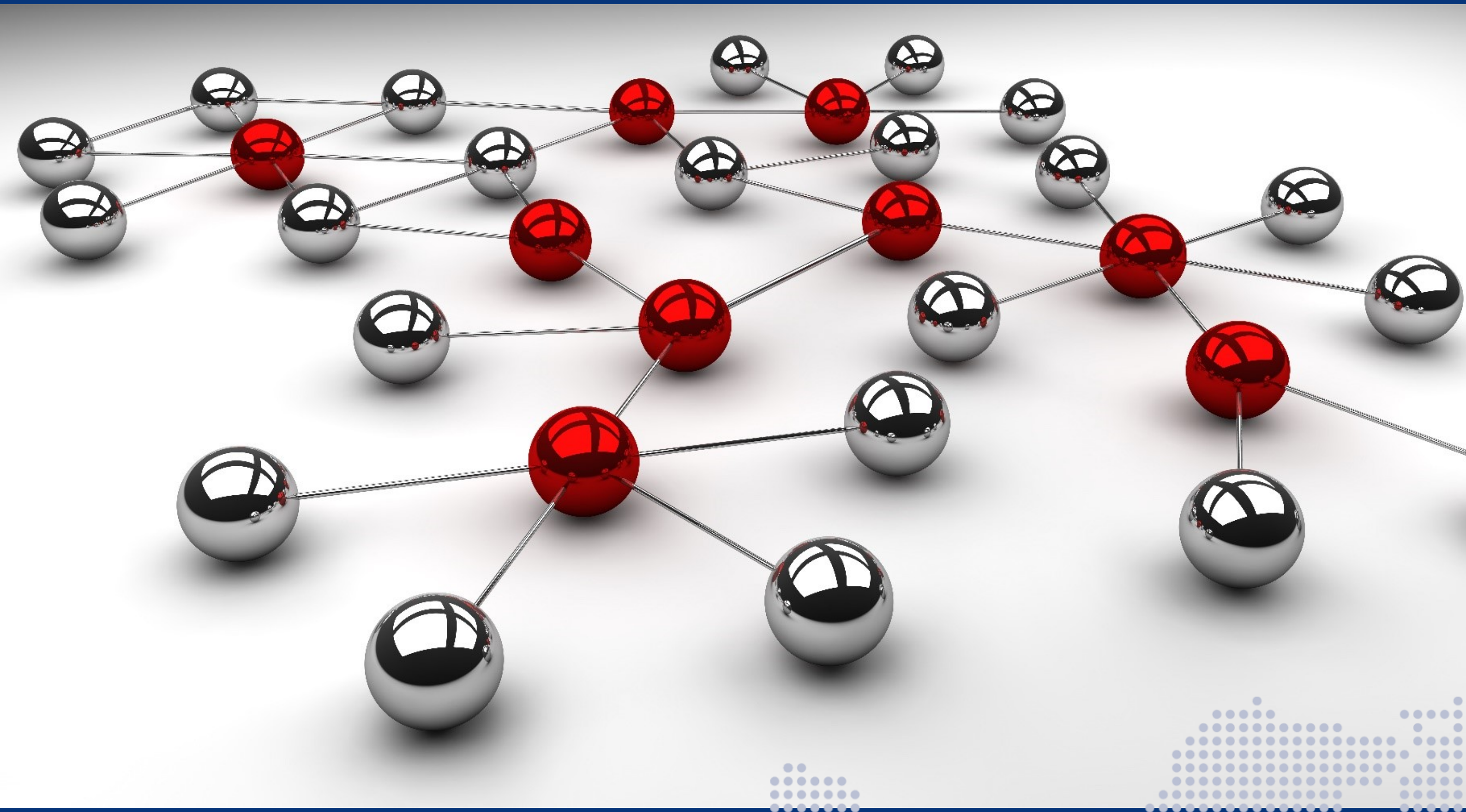
# Machine-learning classification of minimum wage workers

## Final report



LE London Economics

## About London Economics

London Economics is one of Europe's leading specialist economics and policy consultancies and has its head office in London.

We advise clients in both the public and private sectors on economic and financial analysis, policy development and evaluation, business strategy, and regulatory and competition policy. Our consultants are highly-qualified economists with experience in applying a wide variety of analytical techniques to assist our work, including cost-benefit analysis, multi-criteria analysis, policy simulation, scenario building, statistical analysis and mathematical modelling. We are also experienced in using a wide range of data collection techniques including literature reviews, survey questionnaires, interviews and focus groups.

**Head Office:** Somerset House, New Wing, Strand, London, WC2R 1LA, United Kingdom.

w: londoneconomics.co.uk e: info@londoneconomics.co.uk 🐦: @LondonEconomics
t: +44 (0)20 3701 7700 f: +44 (0)20 3701 7701

## Acknowledgements

We would like to acknowledge the useful guidance and feedback provided by the Low Pay Commission throughout this research, in particular Ronia Salman and Tim Butcher. We would also like to thank the ONS for access to data sources and advice relating to the necessary data and data availability. This report also benefited from the insights provided by the participants at the Low Pay Commission workshops on 1st September 2022, 27th April 2023, and 7th September 2023, with particularly helpful feedback from the discussants Mirko Draca and David Zentler-Munro. Responsibility for the contents of this report remains with London Economics.

## Authors

**Su-Min Lee** Senior Economic Consultant

**James Forrester** Economic Consultant

**Lucy Manly** Economic Consultant

**Jenny Liu** Research Intern

**Gavan Conlon** Partner

Wherever possible London Economics uses paper sourced from sustainably managed forests using production processes that meet the EU eco-label requirements.

London Economics Ltd is a Limited Company registered in England and Wales with registered number 04083204 and registered offices at Somerset House, New Wing, Strand, London WC2R 1LA. London Economics Ltd's registration number for Value Added Tax in the United Kingdom is GB769529863.

# Table of Contents

Page

# Executive Summary

London Economics have undertaken research exploring a range of machine-learning techniques to understand

1) which groups of workers are more likely to be minimum wage workers, and
2) key groups of workers and minimum wage workers.

It is important to understand which types of workers are more likely to be minimum wage workers and how minimum wage workers differ as it provides policymakers with an understanding of how changes to the minimum wage may impact different types of workers.

While there are many different ways to segment the labour force, the machine-learning classification methods used in this analysis provide an  objective characterisation of groups within the labour force and among minimum wage workers. These methods have not been widely used in minimum wage research, and so they may offer novel insights into the low-paid labour market and allow us to more precisely define groups affected by the minimum wage.

## Methods

A range of machine-learning classification methods are used in the analysis to categorise different types of workers. These methods can be grouped into two different types of methods.

The first group is made up of supervised tree-based classification methods. A classification method is ‘supervised’ when it uses characteristics to predict an explicitly chosen outcome. The outcome in this context is whether a worker is a minimum wage worker[1].

The tree-based methods (decision tree, random forest, and boosted gradient trees) split the sample into different groups that share characteristics (for example, ‘workers under the age of 25 working in manufacturing and health services’).

These groups are formed to separate those who are minimum wage workers and those who are not as much as possible using different combinations of characteristics. This provides an understanding of which characteristics are most important in predicting whether a worker is a minimum wage worker or not. A single decision tree splits the sample by different characteristics to form one set of groups. Ensemble methods (random forest and boosted gradient trees) combine multiple decision trees to form more robust conclusions.

The second method is the Latent Dirichlet Allocation, which is an unsupervised classification method. A classification method is ‘unsupervised’ if the groups are formed without predicting a particular outcome. The groups are formed by characteristics that often appear together (for example, working in professional occupations being strongly correlated with having a university degree). The Latent Dirichlet Allocation is a ‘probabilistic’ model, so it estimates the probability that each worker is a member of each group based on their characteristics.

[1] For the main analysis this is defined as the worker having an hourly pay of five pence above their relevant minimum wage or less.

### Data

The analysis uses information from the Labour Force Survey from 2013 to 2022, excluding quarters when hourly pay information may have been distorted by the furlough scheme. The Labour Force Survey is used as it includes a wide range of personal characteristics: age, sex, ethnicity, region of residence, region of work, highest educational qualifications, occupation of work, industry of work, disability status, marital status, and number of dependent children.

### Results

The decision tree analysis identifies fifteen groups of workers, each defined by a combination of characteristics. For example, the group of workers that has the highest concentration of minimum wage workers are workers aged 21 and over working in sales and elementary administrative/service occupations in the accommodation and food, education, and arts industries. Over half of that group (52.4%) are minimum wage workers, between three and four times the percentage of workers across the whole labour force who are minimum wage workers. Occupation is the most important characteristic in predicting whether a worker is a minimum wage worker in the decision tree, followed by industry of work.

Random forests and gradient boosted trees are used to test how robust the findings from the decision tree are. The decision tree is estimated using one subset of the available data[2], so conclusions made from a single decision tree may change depending on which subset of the available data is used. Random forests and gradient boosted trees avoid this concern by combining many different decision trees. The conclusions made from the decision tree can also be made from the random forest and gradient boosted trees. For example, occupation is by far the most important characteristic that predicts whether a worker is a minimum wage worker, followed by industry of work and highest educational qualification achieved.

We also test tree-based models excluding job characteristics (such as occupation and industry). This is an important extension, as it could allow us to predict potential minimum wage workers amongst the unemployed or inactive. However – as might be expected given the importance of job characteristics in the initial analysis – we find that these models do not perform as well. This suggests that personal characteristics alone – at least those that can be measured in the available data – are not the key determinants of whether a worker is a minimum wage worker.

The Latent Dirichlet Allocation identifies ten groups of workers based on clusters of characteristics. Although this method does not identify groups based on whether they are minimum wage workers or not, the Latent Dirichlet Allocation identifies three clusters of characteristics that have a much higher concentration of minimum wage workers than the other seven.

The characteristics that appear disproportionately often within these three clusters can be interpreted broadly as

1) older workers with lower-level qualifications in elementary occupations,
2) workers in education and health and social work activities with vocational backgrounds who are disproportionately likely to be from ethnic minority backgrounds, and
3) younger workers with Level 3 academic qualifications (e.g., A-Level) as their highest qualification working in sales and elementary occupations.

[2] The remaining part of the data is used to evaluate the predictions made by the model.

It is important to note that given the probabilistic nature of the Latent Dirichlet Allocation, these clusters should not be interpreted as mutually exclusive groups of workers but of combinations of characteristics.

**Evaluation**

The evaluation of the tree-based methods that aim to predict whether a worker is a minimum wage worker can be based on how often they correctly classify workers as being minimum wage workers[3], known as precision. A higher precision is the equivalent of a lower proportion of false positives, the proportion of workers who are mistakenly classified as being minimum wage workers.

In addition, the evaluation can be based on the proportion of all minimum wage workers that the prediction model identifies as minimum wage. A higher recall is the equivalent of a lower proportion of false negatives, the proportion of minimum wage workers who are mistakenly not classified as being minimum wage workers.

Predictions based on the decision tree and random forest perform reasonably similar to predictions based on regression models (based on their precision and recall), while the gradient boosted trees perform better than those alternatives.

The Latent Dirichlet Allocation does not explicitly predict whether workers are minimum wage workers, so the evaluation of the Latent Dirichlet Allocation is primarily conducted by testing how much the groups change when small changes are made to the setup of the analysis. A range of robustness checks are implemented which suggest that the characteristics of the groups identified are generally robust to changes in how the Latent Dirichlet Allocation is modelled.

**Limitations and caveats**

The most important limitations and caveats to note when implementing machine-learning classification in this context are

- the interpretation of the results of machine-learning classification can often be complex,
- the models predict outcomes or identify correlations between characteristics and do not necessarily provide causal links between characteristics and whether a worker is a minimum wage worker (for example, it may not be appropriate to extend conclusions based on employees to those who are not in work), and
- the conclusions made from the models are limited by the information available from the data which can suffer from issues such as measurement error or having relatively few variables available to help predict whether a worker is a minimum wage worker.

[3] Where classifying a worker as a minimum wage worker is based on whether the probability that they are a minimum wage worker, based on their characteristics, is greater than a given threshold (e.g., 30%).

# 1 Introduction

Understanding the impact of the minimum wage on labour market outcomes is a key component of minimum wage policymaking. Existing research suggests that there has been a limited impact on aggregate labour market outcomes such as employment.

However, the impact of changes to the minimum wage may differ across different types of workers, and different types of low pay workers. Younger low-paid workers may respond differently to changes in the minimum wage to those who are older, while low-paid workers who have dependent children may respond differently to those who do not.

There are many ways that workers may be different, so in this report, **machine-learning classification methods are used to identify salient combinations of characteristics from a wide range of characteristics in a systematic and objective manner**.

Two machine-learning classification methods are used: **tree-based** classification and **Latent Dirichlet Allocation (LDA)** classification.

- **Tree-based classification methods such as decision tree/random forest/boosted gradient trees** classification are **supervised** learning techniques, where supervised techniques explicitly use characteristics to predict an outcome – in this case whether a worker is a minimum wage worker or not.
- In contrast, LDA classification is an **unsupervised** learning technique, where unsupervised techniques do not predict an outcome, but in this case cluster groups of workers together based on shared characteristics (without reference to the outcome variable, minimum wage status).

We apply these methods using information from the **Labour Force Survey** which includes a wide range of personal characteristics to identify groups of workers using a sample from 2013 to 2022.

Despite differences in their approach, **both classification methods are able to identify groups with high concentration of minimum wage workers and identify a variety of types of minimum wage workers**. Tree-based methods are able to predict which workers are minimum wage workers as well or better than traditional regression methods, as long as job variables are included in the analysis.

# 2 Contributions to related literature

The methods used in this analysis contribute to previous research through **identifying groups of workers (and minimum wage workers) using a systematic and objective approach that incorporates a wide range of characteristics**.

These **groups can be used to investigate how the impact of labour market interventions, such as the minimum wage, differ across salient groups**. In addition, the use of machine-learning classification can **provide potential treatment and control groups** to evaluate impact on labour market outcomes[4].

The contributions of both machine-learning classification methods (although they are significantly different) are that they:

- **prioritise the most important variables** that mark distinctions between different groups of workers (between different types of workers or between minimum and non-minimum wage workers),
- can explore **complex interactions** between variables as well as **non-linear/non-parametric relationships within a variable** (such as age) compared to other classifications, such as Principal Component Analysis and Factor Analysis, which do not,

Our use of supervised machine learning methods is similar to the methods used by of Cengiz et al. (2022) who use tree-based methods to predict whether an individual is a minimum wage worker or not based on personal characteristics.

We complement this methodology by also applying an unsupervised machine learning method (Latent Dirichlet Allocation) that agnostically identifies groups of workers with shared characteristics that often coincide, without ex-ante predicting whether they are a minimum wage worker or not. We can then assess the different concentration of minimum wage workers within these groups.

## 2.1 Aggregate impact of the minimum wage in the existing literature

Much of the main research in the UK context has focused on the introduction of the National Minimum Wage (NMW) in 1999; the introduction of the National Living Wage (NLW) in 2016; changes in age eligibility for minimum wages; and subsequent upratings of the minimum wages. These studies predominantly conducted difference-in-differences analyses, comparing outcome changes in the 'treatment group' (affected workers) following the intervention to outcome changes in a suitably chosen 'control group' of workers (such as those in receipt of marginally higher wages), or comparing across low- and high-wage regions, demographic groups or firms (Dube, 2019). While the minimum wage does not vary across regions and demographic groups, the impact might vary significantly.

A meta-regression analysis by Rand Europe (Hafner et al., 2017) found no overall 'genuine' adverse impact effect of the NMW, whether on employment, hours, or employment retention probabilities. They note that the majority of existing empirical primary studies did not pay specific attention to

[4] The decision tree and random forest analysis provides an estimate of the probability that a worker is a minimum wage worker based on their characteristics. Changes in labour market outcomes of those with similar probabilities but different minimum wage statuses (one is earning the minimum wage and another above the minimum wage) could be compared to estimate the impact of changes to the minimum wage on labour market outcomes.

potentially particularly vulnerable labour market groups. A study by Manning (Manning, 2016) controlled for gender, age, and region effects in the UK, but still found no clear effect on employment within the first ten years after the introduction of the NMW.

Similarly, more recent analyses of the introduction of the NLW (Aitken et al., 2019; Cribb et al., 2021; Datta et al., 2021; Dube, 2019), using a range of different methodologies, did not find a substantial effect the NLW on low-wage employment in the UK.

**However, focusing on the aggregate effect may mask the impact on specific subgroups**. A significant share of workers earn significantly above the minimum wage and are not likely to be impacted by changes to the minimum wage. The share of workers who were earning around the level of the minimum wage varies greatly across regional and demographic groups, as well as across industries and job types. Given this, many US studies in particular have focused on subgroups based on demographic group or industry, such as **teenagers or restaurant workers**, as there is a larger share of these groups of workers who are earning near the minimum wage compared to the entire population. Adopting this approach would make it easier to detect any potential effects of changes in the minimum wage (Dube, 2019). Some UK research has studied the effects on gender and/or age, and there were a range of studies in the early 2000s that studied specific industries likely to be affected by the NMW, as will be described overleaf.

## 2.2 Impact by group in the existing literature

Some studies have found negative effects on the employment opportunities of particular groups in some time periods under certain model specifications:

- **part-time workers** (Hafner et al., 2017) and in particular **female part-time workers** (Aitken et al., 2019; Dickens et al., 2015),
- **care home workers** (Georgiadis, 2006; Machin et al., 2002; Vadean & Allan, 2021; Wilson & Machin, 2004),
- **service industry employees** (Fidrmuc & Tena, 2013), and
- **low-skilled low-wage workers holding automatable jobs** (Lordan, 2019).

Studies evaluating the impact by either gender or age alone (Bryan et al., 2013; Dickens & Draca, 2005; Stewart, 2004) have not found any significant and robust impact of the minimum wage, while a recent study by London Economics (Conlon et al., 2023) found no negative aggregate employment effects for young people in the aftermath of the reduction of the age of entitlement to the National Living Wage in 2021. Datta, Machin, and McKnight (2021) found that the introduction of the NLW actually had a positive effect on employee retention for **women** and **workers with disabilities**. Aitken, Dolton and Riley (2019) found no statistically significant employment effects on workers of **Pakistani or Bangladeshi origin**, who have the highest coverage rates, although they found a statistically significant effect on workers of **Indian origin**. Other industry case studies, such as in **social care** (Georgiadis, 2021; Giupponi & Machin, 2018), **hairdressing** (Druker et al., 2002), **textiles** (Heyes & Gray, 2001; Lucas & Langlois, 2003), **hospitality** (Lucas & Langlois, 2003; Norris et al., 2003) and **horse racing** (Winters, 2001) generally found little effect. Two studies (Dickens et al., 2009; Stops et al., 2012) compared lower wage with higher wage areas in the UK, as the introduction of the NMW affected them differently, but did not find any substantial change in employment rates. Dickens and Lind (2018) also found no significant change in employment rates in response to the NLW.

There has been some recent work that investigates potential labour market impacts across multiple interactions of characteristics.

Dube (2019) constructed 96 groups defined by 12 regions, 4 age categories and 2 gender categories, and found that the affected share (defined as the share of workers earning below the 2018 NLW) varied between 6% and 34%. Butcher and Dickens (2022) constructed 320 groups from the interaction of 20 regions, eight age groups, and two gender groups. They used variation across groups in their exposure through the minimum wage bite and coverage of the minimum wage in 2015. They found no significant negative impacts on employment or hours across the sample time period between 2016 and 2022.

# 3 Data

This section firstly outlines the data source (the Labour Force Survey) used by machine-learning classification methods in this analysis, as well its limitations. This is followed by a discussion of other data sources considered and their limitations.

## 3.1 Data sources

The two types of machine-learning classification methods are explained in the following sections, and both use information from the **Labour Force Survey (LFS)**. This analysis uses LFS data pooled from 2013 to 2022. However, given limitations in measuring hourly pay during the government's Coronavirus Job Retention Scheme (as well as broader data collection issues), **data from the second quarter of 2020 (close to the start of the first lockdown in the UK) until the third quarter of 2021 (after which the furlough scheme had ended) has been excluded**.

The LFS is the largest household survey in the UK recording labour market outcomes and it is used by the Office for National Statistics to construct the official measures of employment and unemployment. The LFS has typically around 80,000-90,000 respondents per quarter (although the number of respondents declined significantly during the period affected by COVID-19) and is a representative sample of the UK resident population, with each household in the survey tracked over five consecutive quarters.

**The main advantage of the LFS is its wealth of personal and labour market characteristics.** The variables that have been used in the analysis are:

- age,
- sex,
- ethnicity,
- region of residence,
- region of work,
- educational qualifications,
- occupation of work,
- industry of work,
- disability status,
- marital status, and
- dependent children.

These characteristics were chosen as the set of characteristics that maximised the sample size. The set of characteristics included in the analysis should ideally not be filtered before the analysis, as an objective method would allow the machine-learning classification methods to indicate which characteristics are salient or not.  However, compared to the chosen characteristics listed above, relatively few individuals provided information about other characteristics. As a result, the inclusion of other characteristics would have resulted in a significant reduction in the sample size.

The time range of almost ten years provides a large sample to train both classifications to investigate long-standing labour market patterns (so that the results are not potentially as sensitive to year-on-year changes). The time range also covers significant increases in the minimum wage which may have changed the types of workers covered by the minimum wage. However, besides the introduction of the National Living wage in 2016 the proportion of workers paid the minimum wage has been relatively stable, suggesting that the types of workers covered by the minimum wage may not have changed significantly over the time range of the sample.

The **main analysis focuses on those who are employees** for two reasons. Firstly, hourly pay information is not available to those who are unemployed or economically inactive. Secondly, the minimum wage does not cover those who are self-employed. However, some tree-based classification analysis is undertaken without job characteristics in Section 6. This analysis is undertaken to understand groups who are unemployed or inactive who may be more likely to be minimum wage workers than other groups if they were employed.

Finally, it should be noted that the **full-time/part-time status of a worker is not included as a characteristic**. Full-time/part-time work is likely to be an important predictor of whether a worker is a minimum wage worker or not, as part-time work is more likely to be paid at the minimum wage than full-time work. However, for the purposes of the following analysis, full-time/part-time status is excluded as a labour market outcome rather than a 'characteristic' of the individual worker.

## 3.2 Limitations of the Labour Force Survey

While the LFS provides a wide range of individual characteristics which will provide useful information in the machine-learning classification, there are two key issues associated with using the LFS (as opposed to other sources such as the Annual Survey of Hours and Earnings) for this analysis:

4) **declining response rates** and
5) **reliability of pay data**.

First, there have been declining response rates in the past 10 years, particularly during the COVID-19 pandemic in 2020, with the total number of observations per quarter falling from 100,000 to 70,000 in some quarters. Information on wages is only collected in Wave 1 and Wave 5 and only from individuals in employment (excluding those self-employed), meaning that wage data are available for around 11%-12% of respondents in a given quarter. Limited sample sizes would lower the statistical power of our analysis. To increase our sample size, **we undertake the analysis across multiple years rather than individual years**.

Second, the data on pay tends to be less reliable than other surveys (such as the Annual Survey of Hours and Earnings), as responses are **self-reported or provided by a proxy respondent** (rather than an employer using payroll records).

The *hour pay* variable is **derived from the gross weekly pay and the usual hours worked reported by the respondent**. This derived variable tends to be lower than the derived hourly pay rate in ASHE, as individuals may not refer to their employment records while completing the survey and incorrectly report their weekly pay and hours worked (unlike employers filling out ASHE using company records). In particular, respondents to the LFS tend to include their unpaid hours whereas these are excluded in ASHE (Low Pay Commission, 2021). This measurement error could lead to issues in the analysis, as discussed further in Section 4.

While analysis of **imputed hourly pay rates could be used**, we propose to use the derived *hourpay* variable. Imputed hourly pay rates have been estimated using the LFS to better match the Annual

Survey of Hours and Earnings[5], the recognised source of estimates of UK pay. However, the findings of the machine-learning classification are likely to be sensitive to the methods used to impute hourly pay rates.

## 3.3 Alternative data sources

**Alternative data sources** were also considered, but limitations of these alternative sources contributed to the choice of the LFS as the preferred suitable data source.

The **Annual Survey of Hours and Earnings (ASHE)** is an annual survey carried out in April each year and based on a 1% sample of employee jobs taken from HM Revenue & Customs (HMRC) PAYE records. The main benefit of ASHE is the larger sample size compared to the LFS and the fact that it is based on employer records. It provides information on earnings, hours of work, as well as a range of other pay information. However, there are few personal characteristics included, which would have significantly constrained the machine learning classification, especially in comparison to the LFS. We therefore decided against using ASHE for the analysis.

Other data that could be considered is **ASHE – 2011 Census Dataset from the Wage and Employment Dynamics (WED) project**. This data is part of a broader WED project that combines existing data to improve labour market data[6]. The ASHE – 2011 Census Dataset combines the high-quality hourly wage data from ASHE with the wide range of personal characteristics from the 2011 Census. A set of variables (age, sex, surname, initials, home and work postcodes, occupation, and industry sector) are used to match individuals' records in ASHE. It enhances ASHE data by merging 317 variables from the 2011 Census. These variables include characteristics such as ethnicity, English language proficiency, country of birth, religion, marital status, health and disability, and educational and vocational qualifications.

Despite the wide range of characteristics available, there are two main limitations in the use of ASHE – 2011 Census data:

- The **most recent year available in the current edition of the data is 2018**. Although the methodology aims to provide insight into the characteristics of minimum wage workers across a reasonable time range, the preferred dataset would ideally include the most recent years.
- **Data linking between ASHE and the 2011 Census is imperfect** and significantly less perfect the further away from 2011. The share of ASHE 2011 records linked to the 2011 Census is just above 60% and drops to around 40% in 2018. Further, the probability of being linked is not random across characteristics. Linkage rates are lower for older workers, while they are higher for those with higher pay. Those living in London are nine percentage points less likely to be linked.

[5] An example presented by the ONS can be found here.

[6] More information can be found on the WED project website here.

# 4 Limitations of machine-learning classification methods

Tree-based classification methods (decision trees/random forests/gradient boosted trees) were chosen due to their ability to identify non-parametric combinations of characteristics that may be important in explaining whether a worker is a minimum wage worker. Further, the tree-based classification methods are complementary: decision trees provide intuitive results where groups can be presented clearly, while random forests and gradient boosted trees provide a robustness check by combining the results of multiple decision trees. The Latent Dirichlet Allocation (LDA) was chosen to provide a holistic overview of groupings within the labour market (without predicting whether a worker is a minimum wage worker): which workers are similar across multiple characteristics, even if those characteristics are not important in predicting whether a worker is a minimum wage worker?

While subsequent chapters discuss limitations in the use of specific methods, this section explains broader limitations of using machine-learning classification to understand minimum wage workers.

One limitation is the **potentially complex results** that machine learning classifications can produce that are difficult to interpret. The interpretability of identified groups may be limited if they are defined by a complex interaction or cluster of characteristics (for both decision trees and the LDA classification).

As a predictive method, both classification techniques **do not provide explicit explanations as to why** a particular characteristic, or cluster of characteristics, may be associated with being a minimum wage worker. This places greater emphasis on using descriptive statistics and existing research to help explain results that may be initially counter-intuitive.

This is especially true of an unsupervised method such as the Latent Dirichlet Allocation. There is no ex-ante guarantee that there would be a sharp definition between groups in the proportion of members who are minimum wage workers.

This merits the use of two different types of methods, such as the decision tree method that focuses on explicitly predicting which groups are likely to contain many minimum wage workers.

Issues of **underfitting** (where the model is too simplistic and averages over important differences between groups) and **overfitting** (where the model is too complex given the data where its findings are informed by noise rather than systematic differences) are discussed in the robustness check section above, although tests have been run to estimate the optimal sensitivity (decision tree) and the optimal number of groups (LDA classification).

The classification is also **limited by the available data**. The classification technique will not be as accurate if there is a key unobservable factor (such as some personality characteristic) that is not available in the data.

**These classification techniques are primarily a predictive and descriptive exercise rather than seeking to identify causal links between characteristics and earnings**. As a result, if these unobserved factors are correlated with some observable characteristics, then the omission of the unobserved factors is not as critical to the performance of the classification. For example, the trait of being able to learn new skills quickly may not be observed directly in the data but will be correlated to characteristics available in the data, such as educational attainment.

The classification may be impacted by **measurement error**, such as when regressing the outcome of interest (whether a worker is a minimum wage worker based on hourly earnings) on dummy variables that indicate whether the worker is a member of each group defined by the machine-learning classification. If there is measurement error in the outcome variable (or the hourly earnings used to derive the outcome variable), this will increase the estimated standard errors of the coefficients of the dummy variables. Further, there may be attenuation bias (i.e., downwards bias in the magnitude of the coefficients) if there is measurement error in the characteristics. Both may reduce the estimated effectiveness of the classifications in predicting whether a worker is a minimum wage worker or not, although it is likely that measurement error impacts hourly earnings information to a greater extent than characteristics such as age.

Further, there are **ethical considerations**. There are several ways in which the results may suffer from bias, in particular when using complex methods such as machine-learning classification[7]. Transparency and explicability may be obscured using more complex methods, which is why we propose to present a variety of different descriptive statistics to explain the results. This will also reduce prejudicial bias when interpreting the results in providing an understanding of the results. The ONS provides a sample representative of the wider UK population in the Labour Force Survey, which reduces potential sample or exclusion biases. While our main analysis focuses on those with job characteristics (i.e., those in employment), we also undertake analysis across the wider population, although without being able to use job characteristics. The use of two contrasting machine-learning classification methods and a range of robustness checks may also help to mitigate potential biases – if the results of one classification are very different to the other further investigation to explain these differences would be undertaken.

[7] A wider discussion of its implications can be found here (ONS, 2021).

# 5 Decision Tree analysis

## 5.1 Methodology

A decision tree takes a sample and repeatedly splits it according to certain characteristics, trying to create groups that separate those who are minimum wage workers and those who are not. The result is a set of groups at the end of the tree (leaf nodes). Those in a given group share a common combination of characteristics (for example, women working in sales and elementary administrative and service occupations in the wholesale, retail, administration, support, and social work industries).

**For avoidance of doubt, there is no pre-determined ordering of characteristics, and the decision tree algorithm chooses which characteristics to split the sample by and in which order.**

To begin with, the entire sample is included at a root node, and then a split is chosen according to a characteristic (for example, based on the workers' occupation[8]). For a given stage in splitting the sample, the chosen split is the one that reduces the average entropy of the tree as much as possible. The average entropy of the tree is the weighted average (by population at each node) of the entropies of each individual leaf node. The entropy of a given node is calculated as

$$Entropy = \sum_{i=MW,NMW} -p_i \times \log_2(p_i)$$

where $p_{MW}$ is the proportion of the group at the node who are minimum wage workers, whereas $p_{NMW}$ is the proportion of the group at the node who are not minimum wage. This entropy is maximised if $p_{MW}$ and $p_{NMW}$ are the same, and minimised when either $p_{MW}$ tends to zero or one (and therefore equivalently for $p_{NMW}$). As a result, an equivalent interpretation is that at each step, the decision method **finds the optimal way to split the sample such that it separates those who are minimum wage workers and those who are not as much as possible**.

**This optimal splitting is how the decision tree prioritises which characteristics are most important in predicting whether someone is a minimum wage worker or not.**

When searching for the optimal split, the decision tree considers both:

- **which characteristic to split on** (for example, whether occupation, industry, or gender is the most important variable in determining a worker's minimum wage status), and
- **what split within a characteristic to use** (for example, whether to split the sample on either side of the age of 25 or the age of 21).

This displays one **key advantage of the decision tree classification over other prediction models**, such as OLS and logit/probit models: it can **search for key characteristics/splits within characteristics over many potentially complex combinations of characteristics** in a way that would be laborious in other prediction models (for example, finding exactly the optimal split or splits across age).

In prioritising the most important splits early on, the decision tree can be **pruned** where a split only occurs if it is significant enough. A benefit of this pruning is the avoidance of categorising workers

[8] This analysis uses two-digit occupation codes from the SOC2010 classification. See Table 6 in **Error! Reference source not found.** for f urther details.

into groups that are too narrowly defined to be of any practical purpose and to **avoid overfitting**, where conclusions (differences between groups) arise from noise in the data.

In this analysis, pruning is undertaken by optimising the choice of a sensitivity parameter that determines how many splits the decision tree makes. The sensitivity parameter is chosen to minimise the cross-validation relative error, which is a measure of the accuracy of the decision tree model when the model is applied to new, unseen data.[9]

[9] See Annex **Error! Reference source not found.** for further details on the choice of sensitivity parameter.

## 5.2 Results

### 5.2.1 Example pathway

As an example of a pathway from the decision tree (see Figure 2 for the full tree), Figure 1 illustrates the group in the decision tree with the highest concentration of minimum wage workers (52%). This section follows the pathway that leads to this group. For simplicity, all other decision tree pathways and nodes are omitted in order to focus on the pathway that leads to this group.

At the root node, the most important characteristic in determining minimum wage status is the worker's **occupation**. Workers in occupations denoted by two-digit codes 51, 54, 61, 62, 71, 91, and 92 are more likely to be minimum wage workers – as highlighted above, compared to an average concentration of minimum wage workers of 14% across the sample as a whole, 33% of workers in these occupations are minimum wage workers.

At the level one node (i.e., the child node of the root node), the most important characteristic in determining minimum wage status among the workers within these occupations is the **industry** in which they are working. Workers in industries denoted by SIC section letters A, G, I, N, P, Q, R, S, and T[10] are more likely than workers in other industries to be minimum wage workers.

At the level two node and level three node (i.e. the child nodes of the level one and level two nodes respectively), **occupation** and **industry** are again in turn the most important characteristics in determining these workers' minimum wage status, with firstly workers in occupations with two-digit codes 71 and 92[11], and secondly workers in industries denoted by section letters A, I, P, R, S, and T[12] being selected by the decision tree algorithm. At the level four node (the penultimate node in this pathway), the decision tree algorithm determines that age is the most important factor in determining these workers' minimum wage status. Workers are then split into those aged under 21 and those 21 or older.

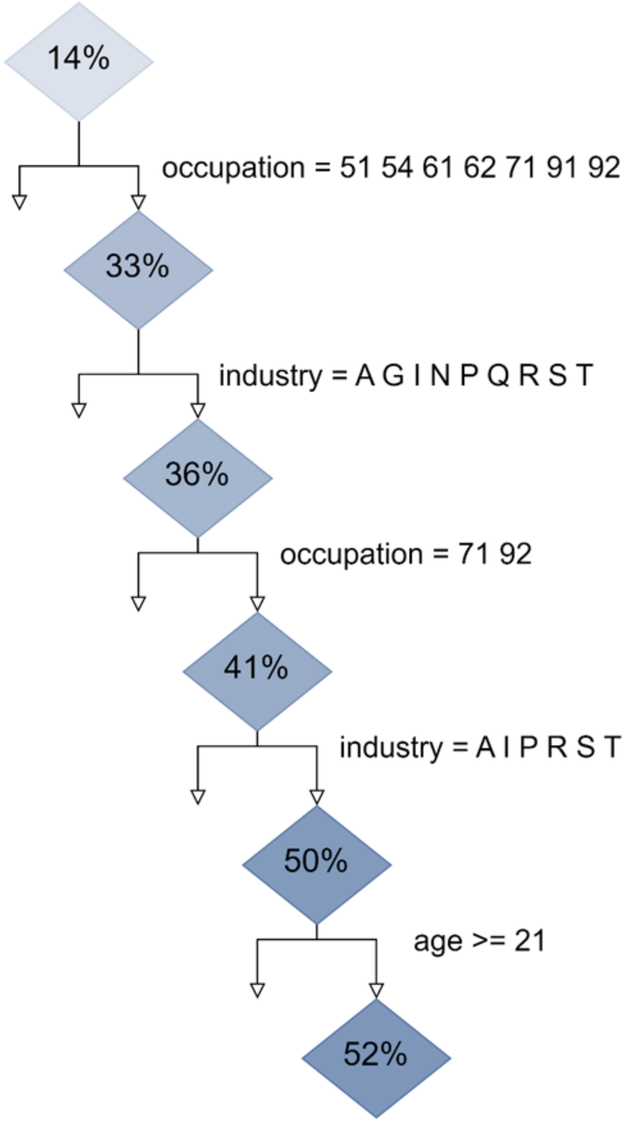
[10] These section letters indicate the industries indicated below. For a full breakdown of the SIC section letters, see Annex A1.1:

- A: Agriculture, Forestry and Fishing
- G: Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles
- I: Accommodation and Food Service Activities
- N: Administrative and Support Service Activities
- P: Education
- Q: Human Health and Social Work Activities
- R: Arts, Entertainment and Recreation
- S: Other Service Activities
- T: Activities of Households as Employers; Undifferentiated Goods- and Services-producing Activities of Households (Own Use)

[11] This denotes workers in sales (71) and elementary administration and service occupations (91). For a full breakdown of the two-digit occupation classification, see Annex A1.2

[12] Denoting workers in Agriculture, Forestry and Fishing (A), Accommodation and Food Service Activities (I), Education (P), Arts, Entertainment and Recreation (R), Other Service Activities (S), and Activities of Households as Employers (T).

Figure 1 Pathway to the group with the highest concentration of minimum wage workers



Note: all other pathways and nodes are excluded for simplicity.

***Source: London Economics analysis of LFS data***

This group (denoted as group 1, given that it contains the largest concentration of minimum wage workers) is therefore made up of workers with the following characteristics:

- Occupation: **sales** (71) and **elementary administration and services** (92)
- Industry: **agriculture** (A), **accommodation and food** (I), **education** (P), **arts** (R), **households** (T) and "**other services**" (S)
- Age: **21 and over**

Over half (52%) of workers in this group are minimum wage workers, meaning that workers in this group are 3.7 times more likely to be minimum wage workers than workers across the whole sample (where the coverage is around 14%).

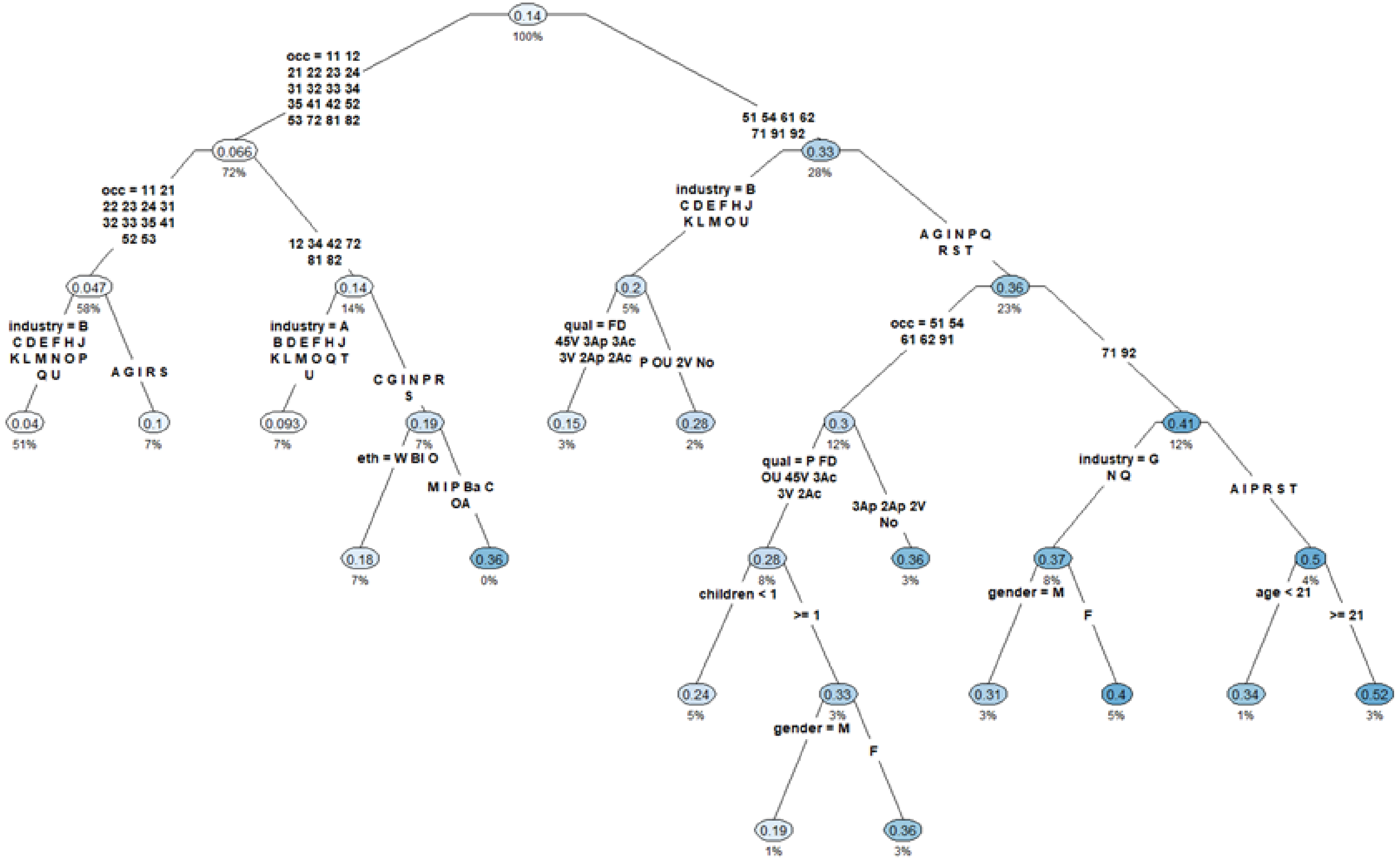
### 5.2.2 Decision tree groups

The decision tree produces **15 groups** of workers at leaf nodes of the tree. The concentration of minimum wage workers within each of these groups ranges from a low of **4%** in the group with the lowest concentration of minimum wage workers, to a high of **52%** in the group with the highest concentration of minimum wage workers.

A worker's occupation is the first and most important determinant of minimum wage status in the decision tree analysis. Subsequent splits introduce other variables such as the industry of work as a key determinant of minimum wage status. Indeed, some of the 15 groups identified by the decision tree (especially those with low a concentration of minimum wage workers) are defined solely by occupation and industry. The full decision tree is presented in Figure 2.

The interpretation of the groups is reasonably intuitive. For example, workers over the age of 21 working in sales and elementary administrative and service occupations in the agriculture, accommodation and food, education, arts, households and "other" industries has been identified as a minimum wage group. On the other hand, workers in the highest skilled occupations working in all industries except agriculture, retail and wholesale, accommodation and food, arts and "other" industries is identified as a group with a low concentration of minimum wage workers (see Table 1).

## Figure 2 Decision tree (main analysis)



Note: The number inside of the node is the proportion of workers at that node who are minimum wage workers, while the percentage underneath is the proportion of all workers in the sample contained within the node.

***Source: London Economics analysis of LFS data***

### 5.2.3 Characterising minimum wage groups

In the context of the decision tree analysis, minimum wage groups are characterised as precise groupings of workers who share common characteristics. Every worker can be explicitly assigned to exactly one group based on their characteristics. The groups can therefore be precisely defined with reference to the shared characteristics of the workers contained within the group.

Table 1 below shows the 15 groups identified by the decision tree, ranked from highest to lowest in order of “precision”, i.e., ranked by the concentration of minimum wage workers contained within the group.

Table 1 Decision tree groups (ranked by precision)

| Rank (by precision) | Precision | Recall | % of sample | Description |
|---|---|---|---|---|
| 1 | 52.4% | 11.9% | 3.2% | Workers aged 21 or over working in sales and elementary administrative and service occupations in the agriculture, accommodation and food, education, arts, households and "other" industries |
| 2 | 40.2% | 14.3% | 5.0% | Women working in sales and elementary admin and service occupations in the wholesale and retail, administration and support, and social work industries |
| 3 | 36.2% | 8.9% | 3.4% | Lower qualified workers in skilled agriculture and other trades, care, leisure, and elementary trades in the wholesale and retail, administration and support, social work, agriculture, accommodation and food, education, arts, households, and "other" industries |
| 4 | 36.0% | 1.1% | 0.4% | Ethnic minority workers (apart from black and "other" ethnicities) in mid-skill level occupation in the manufacturing, wholesale and retail, accommodation and food, administration and support, education, arts and "other" industries |
| 5 | 35.8% | 7.3% | 2.9% | Relatively highly qualified women with children working in skilled agriculture and other trades, care, leisure, and elementary trades in the wholesale and retail, administration and support, social work, agriculture, accommodation and food, education, arts, households, and "other" industries |
| 6 | 34.4% | 1.5% | 0.6% | Workers under the age of 21 working in sales and elementary administrative and service occupations in the agriculture, accommodation and food, education, arts, households, and "other" industries |
| 7 | 30.8% | 6.6% | 3.0% | Men working in sales and elementary administrative and service occupations in the wholesale and retail, administration and support and social work industries |
| 8 | 28.2% | 3.7% | 1.8% | Broadly lower qualified workers and those with a postgraduate working in a range of occupations such as skilled trades, care, leisure, sales, and elementary trades across a range of industries including manufacturing, construction, transport, financial and insurance activities, and real estate |
| 9 | 24.3% | 8.1% | 4.7% | Relatively highly qualified workers without children working in skilled agriculture and other trades, care, leisure, and elementary trades in the wholesale and retail, admin and |

| Rank (by precision) | Precision | Recall | % of sample | Description |
|---|---|---|---|---|
| | | | | support, social work, agriculture, accommodation and food, education, arts, households, and "other" industries |
| 10 | 19.1% | 0.9% | 0.6% | Relatively highly qualified men with children working in skilled agriculture and other trades, care, leisure, and elementary trades in the wholesale and retail, administration and support, social work, agriculture, accommodation and food, education, arts, households, and "other" industries |
| 11 | 18.2% | 8.5% | 6.6% | White, black, and "other" ethnicities in mid-skill level occupation in the manufacturing, wholesale and retail, accommodation and food, administration and support, education, arts and "other" industries |
| 12 | 15.1% | 3.1% | 2.9% | Mid-highly qualified graduates working in a range of occupations such as skilled trades, care, leisure, sales, and elementary trades across a range of industries including manufacturing, construction, transport, financial and insurance activities, and real estate |
| 13 | 10.1% | 4.9% | 6.8% | Workers in the highest skilled occupations working in agriculture, retail and wholesale, accommodation and food, arts and "other" industries |
| 14 | 9.3% | 4.6% | 6.9% | Workers in mid-skill level occupations working across industries apart from in manufacturing, wholesale and retail, accommodation and food, admin and support, education, arts and "other" industries |
| 15 | 4.0% | 14.5% | 51.2% | Workers in the highest skilled occupations working in all industries except agriculture, retail and wholesale, accommodation and food, arts and "other" industries |

Note: Precision indicates the probability that a randomly chosen member of a group is a minimum wage worker. Recall indicates the proportion of all minimum wage workers that are contained within that group.

***Source: London Economics analysis of LFS data***

By summing the recall of groups 1, 2 and 3 (as ranked by precision), it can be observed that these three groups contain **over a third** of all minimum wage workers, despite only containing **less than 12% of the sample.**

## 5.3 Evaluation

### 5.3.1 Precision and recall

For a group identified as **‘minimum wage group’**, two measures can be calculated that capture two desirable characteristics for the output of the classification:

- **Precision:** what proportion of the identified group are minimum wage workers, and
- **Recall:** what proportion of minimum wage workers are included in the identified group.

Precision and recall measures have been used in a variety of machine learning and economics contexts, such as energy markets forecasting (Shao et al., 2020), predicting credit scoring and bankruptcy likelihood (Boughaci et al., 2021), and measuring investor sentiment from news media (Obaid and Pukthuanthong, 2022).

The “**precision**” with which group 1 allows us to predict the minimum wage status of a worker in this group is therefore 52%, corresponding to the probability that a randomly selected worker from the group is a minimum wage worker.

At the same time, 12% of all minimum wage workers are contained within group 1[13]. That means that the “**recall**” of this group is 12%.

The decision tree was estimated based on a **training dataset** made up of a random selection of 20% of the entire sample of the pooled LFS data. Based on the decision tree model, the groups found have been applied to the **testing data** (made up of the remaining 80% of the dataset). The decision tree assigns every individual to a group and a corresponding probability estimate of minimum wage status based on their characteristics.

It is then possible to predict a worker’s minimum wage status according to their characteristics. The prediction rule can be formulated based on a specified (arbitrary) probability threshold, where individuals in groups with a probability above the threshold are predicted to be minimum wage workers. Those in groups with a probability below the threshold are predicted not to be minimum wage workers.

The performance of the decision tree’s classification can then be shown in a **confusion matrix** which presents the number of correct and incorrect predictions made by the classification. These values can be interpreted as measures of Type I errors (false positives) and Type II errors (false negatives).

For example, we can choose a prediction threshold such that it is predicted that all workers contained within groups 1, 2, and 3 are minimum wage workers, and workers in other groups are not minimum wage workers.

This approach is equivalent to assigning a predicted probability threshold of 36% when categorising a worker as minimum wage. That is because these three minimum wage groups all have a precision

[13] In Figure 2, the ‘3%’ figure placed underneath the leaf node containing the group with the highest concentration of minimum wage workers (in the bottom right corner of the diagram) represents the proportion of all workers contained within the group, whereas 12% represents the proportion of all minimum wage workers that are included in the group.

of 36.2% or higher, whereas the group with the fourth highest concentration of minimum wage workers all have a precision of less than 36%.[14]

That is, under a **binary classification with a 36% threshold**, workers who are assigned a probability of more than 36% by the decision tree are predicted to be minimum wage workers. Those assigned a probability less than 36% are not predicted to be minimum wage workers.[15]

Consider the binary confusion matrix based on a **36% prediction threshold** using the testing sample of 147,326 observations. Within each cell, the number in the top row indicates the frequency, whereas the % below indicates the proportion of all observations that fall into that category. The four percentage values therefore sum to 100%.

**Table 2 Binary confusion matrix using a 36% prediction threshold**

| | **The worker is not a minimum wager worker** | **The worker is a minimum wage worker** |
|---|---|---|
| The worker is **not predicted** to be a minimum wage worker | 117,295<br>(79.6%) | 12,825<br>(8.7%) |
| The worker is **predicted** to be a minimum wage worker | 9,886<br>(6.7%) | 7,320<br>(5.0%) |

***Source: London Economics analysis of LFS data***

As **precision** is defined as the proportion of workers predicted to be minimum wage workers who are in fact minimum wage workers, it is possible to calculate the precision of this prediction rule by dividing across the bottom row. The **precision** at this threshold is therefore **42.5%** $= \frac{\mathbf{7{,}320}}{\mathbf{9{,}886+7{,}320}}$.

As **recall** is the proportion of all minimum wage workers who are predicted to be minimum wage workers, it is possible to calculate the recall of the prediction rule by dividing across the right-hand column. The **recall** at this threshold is **36.3%** $= \frac{\mathbf{7{,}320}}{\mathbf{12{,}825+7{,}320}}$.

### 5.3.2 Precision-recall curves

**One limitation of this evaluation method is the necessity to choose this probability threshold**, over which it is assumed that the prediction model predicts that a worker is a minimum wage worker. To some extent, as highlighted above, **the choice of the 36% threshold is arbitrary.** To put this threshold into context, Cengiz et al. (2022) define a ‘high-probability group’ of workers using 39% probability threshold to include individuals with the 10% highest probability of being a minimum wage worker (according to their preferred boosted tree model).

Instead of making an arbitrary choice to create one confusion matrix, it is possible to plot the precision and recall for a range of thresholds (i.e., a range of decision rules defined by a probability threshold).

Precision-recall curves[16] can plot the performance of the different methods by tracking precision and recall across different classification parameters. In this case, the parameter used is the probability threshold used to determine whether a worker (given the estimated probability of being

[14] The precision of group 4 rounds to 36.0% but in fact sits below 36.0%.

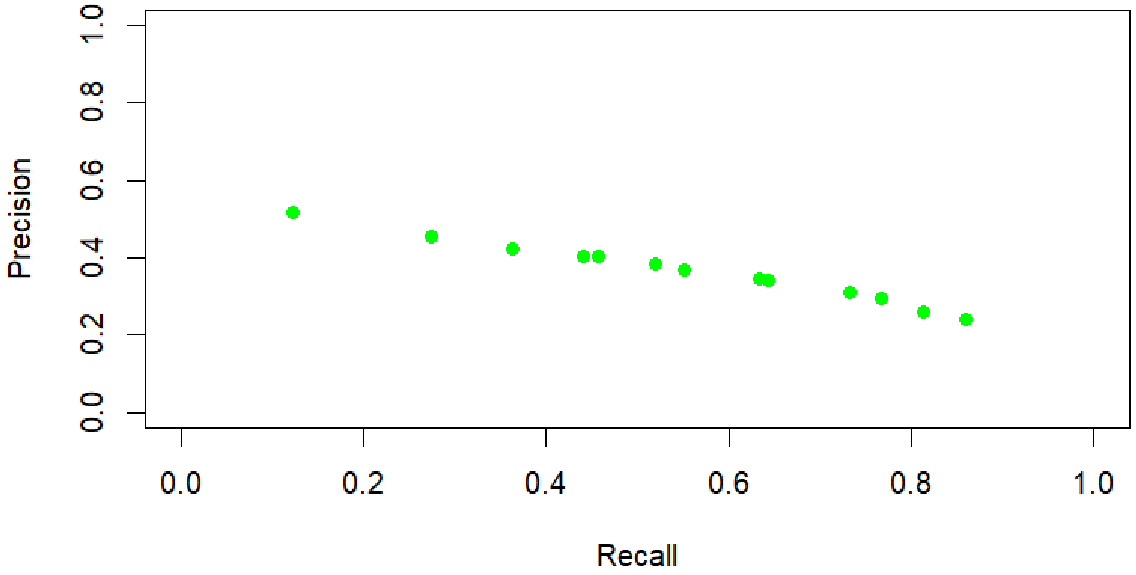
[15] Given the discrete nature of the groups and the resulting discrete probabilities assigned to all workers in the testing data, any threshold within the range of 36.0% and 36.2% would be equivalent. 36.0% is chosen as an illustrative example.

[16] Also known as receiver operating characteristic (ROC) curves

a minimum wage worker) is predicted to be a minimum wage worker according to the prediction model. The further to the top-right hand corner the precision-recall curve is, the better performing the classification is (i.e., one method dominates another if the precision-recall curve has a better precision for every given recall, or vice versa).

Figure 3 shows the precision-recall curve for the decision tree method across a range of thresholds. Each point on the diagram indicates a specified probability threshold used to determine whether a worker (given the estimated probability of being a minimum wage worker) is predicted to be a minimum wage worker according to the prediction model. The threshold in question is indicated by the label next to each point on the diagram.[17]

**Figure 3 Precision-recall curve for the decision tree**



Note: Precision indicates the probability that a randomly chosen member of a group is a minimum wage worker. Recall indicates the proportion of all minimum wage workers that are contained within that group. Each numerical label specifies a probability threshold that is used to determine whether a worker is predicted to be a minimum wage worker or not.

***Source: London Economics analysis of LFS data***

### 5.3.3 Comparison with alternative methods

The predictive performance of the decision tree method has been evaluated relative to different predictive models. Specifically, this has been tested using precision and recall measures on regression methods. However, it should be noted that the additional value of the machine-learning classifications to minimum wage research does not rely solely on its ability to predict whether an individual is a minimum wage worker.

In other words, this aspect of the evaluation of the decision tree method therefore asks to what extent these group fixed effects have greater explanatory power compared to other specifications when predicting whether an individual is likely to be a minimum wage worker or not?

**Four alternative methods have been used as a benchmark for the decision tree classification**. The first three methods use the same specification, namely the two continuous variables (age and

[17] Some labels are omitted to avoid overlapping and preserve presentability.

number of children[18]) and dummies for every unique value of the 10 remaining categorical variables. This specification is run for the following three methods:

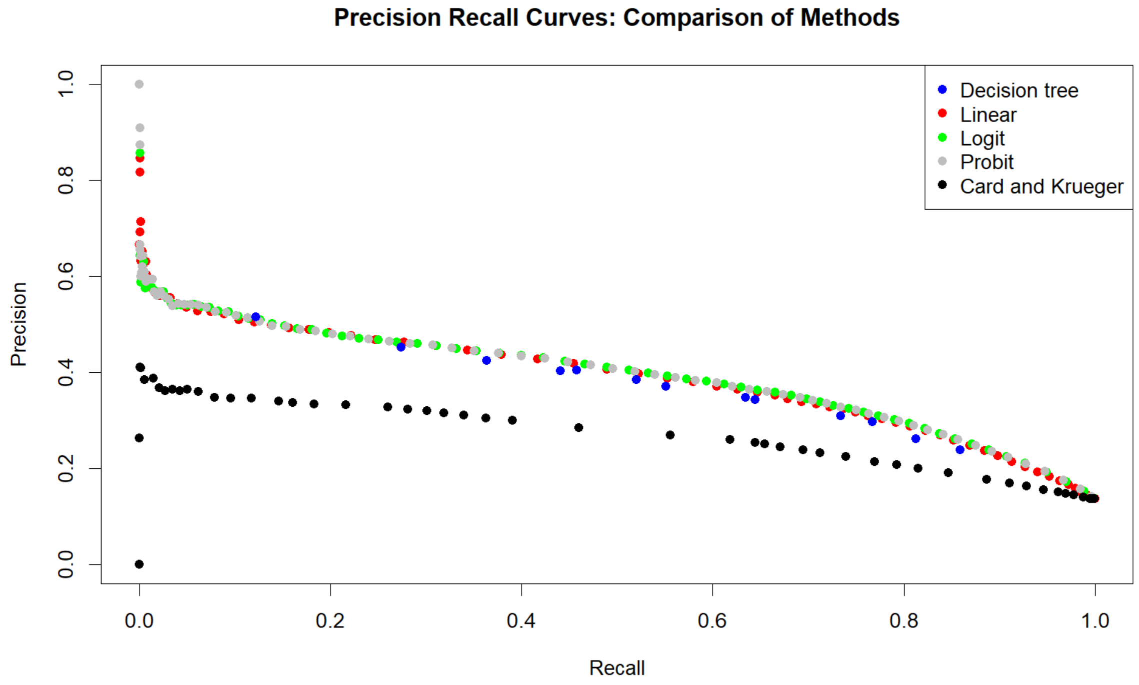
- Linear probability model (OLS with dummy variables for each characteristic)
- Logit model
- Probit model

A fourth method uses a variation on the linear probability specification adopted by Card & Krueger (1995). This is the reference model used by Cengiz et al. (2022). The specification has been adapted to the data available in the LFS and to suit the UK (rather than US[19]) context:

- Three-way interaction between **young adult** (**age 18-25**), **non-white ethnicity** dummy variable, and **gender**
- Three-way interaction between **age**, **qualification level**, and **gender**
- **Quadratic** and **cubic** terms of the **age** variable
- **Non-white ethnicity** dummy variable

Figure 4 presents the precision and recall values for each of these methods across a range of probability thresholds, in a similar manner to Figure 3.

### Figure 4 Comparison of precision-recall evaluation for different methods



Note: Precision indicates the probability that a randomly chosen member of a group is a minimum wage worker. Recall indicates the proportion of all minimum wage workers that are contained within that group. Each point represents a different probability threshold that is used to define whether a worker is predicted, based on their estimated probability, of being a minimum wage worker.

***Source: London Economics analysis of LFS data***

[18] For the purposes of this evaluation, it is assumed that individuals with four or more children in the LFS dataset have four children

[19] For example, the dummy for Hispanic workers used by Card and Krueger (1995) has not been included in this specification.

The further to the top-right hand corner the precision-recall curve is, the better performing the classification is (i.e., one method dominates another if the precision-recall curve has a better precision for every given recall, or vice versa). The plots of each alternative method can then be compared.

Predictions based on the decision tree show **substantially greater predictive power than the Card and Krueger specification**, while the decision tree predictions show similar (albeit slightly worse) predictive power than the **linear probability model**, **logit** and **probit** models using dummies.

It should be noted that the decision tree model is attempting to make predictions based on **15 groups**, rather than almost **100 variables** in the dummy specification. Furthermore, the added value of the decision tree is not just to predict who is likely to be a minimum wage worker, but also **characterising different kinds of minimum wage workers**.

## 5.4 Sensitivity and robustness checks

### 5.4.1 Sensitivity to changing the probability threshold to be classified as a minimum wage worker

Changing the **probability threshold requirements to be part of a group**: given that the prediction of a worker's classification as a "minimum wage" worker is probabilistic (rather than deterministic), a range of threshold requirements have been implemented. This has been done using the precision-recall analysis in Section 5.3.2.

### 5.4.2 Altering the definition of a minimum wage worker

In this subsection, decision trees using different definitions of a minimum wage worker are presented. These are compared to the decision presented in the main analysis (Figure 2), where the definition of a minimum wage worker is one earning five pence above the relevant minimum wage and below.

To aid comparisons, splits in the decision tree are highlighted to indicate how they are similar or different to the decision tree presented in the main report:

- Splits highlighted in dark green indicate that the variable used and the split within the variable are exactly the same as in the main decision tree.
- Splits highlighted in light green indicate that the variable used is the same as in the main decision tree, but the precise split within the variable differs to that of the main tree.
- Splits highlighted in orange indicate that the variable used is different to the variables used in the main decision tree.
- Splits highlighted in blue indicate that the split does not occur in the main decision tree.

**Threshold of 110% of the relevant minimum wage (Figure 5)**

Occupation and industry are still the first and most important variables used by the algorithm to split the sample. Where splits are made using occupation and industry, the occupations used to split the sample at root node and second and third tier nodes are the same or very similar (e.g. the initial occupation split is exactly the same, and the subsequent split by industry at the second tier node on the RHS is exactly the same, but the split by occupation at the second tier node on the LHS differs in respect of one occupation category (41 is now grouped with the other occupations with a higher concentration of minimum wage workers).

The group with the highest concentration of minimum wage workers in both cases is characterised as being over the age of 21, in broadly elementary/low skill occupations within broadly similar sectors (although these do differ to some extent between the two decision trees – in the case of the 110% definition, the industries are accommodation/food, admin/support, education and households, whereas the 5p definition excludes admin/support but includes agriculture, the arts and "other" services). In the case of 110% definition, the group with the highest concentration of minimum wage workers is female, whereas both males and females are included in the corresponding group using the 5p definition.

**Threshold of 125% of the relevant minimum wage (Figure 6)**

Using the 125% definition, occupation and industry are still the first and most important variables used by the algorithm to split the sample. The first split differs slightly in terms of which occupations are characterised as being in the more concentrated minimum wage group – more level 2 skills[20] occupations are grouped with the more concentrated minimum wage group when using the 125% definition.

The group with the highest concentration of minimum wage workers using the 125% definition is broadly similar to the group with the highest concentration of minimum wage workers using the 110% definition, differing only in respect of industry (the group under the 125% definition includes a broader group of industries). Both groups contain women over the age of 21 working in sales, elementary trades, elementary administration, and service occupations.

**Threshold consistent with Annual Survey of Hours and Earnings (ASHE) coverage (Figure 7)**

As the minimum wage coverage using LFS data tends to be greater than that when using ASHE, the threshold is lowered and set such that the overall minimum wage coverage is similar to that when using ASHE.

Occupation and industry are again the first and most important variables used by the decision tree to split the sample. There are two groups of minimum wage workers in broadly lower-paying occupations and sectors that are analogous to the groups of high concentration minimum wage workers mentioned from the other decision trees.

However, other groups have a higher concentration of minimum wage workers than this group. Indeed, the group with the highest concentration of minimum wage workers is a very small group of ethnic minority workers who have a broad range of qualifications (ranging from a first degree to no qualifications) living in different parts of the UK (Scotland, the South West, the North West and the East Midlands) who work in a range of sectors and occupations that range from corporate managers to transport and mobile machine drivers and operatives.

[20] As defined by the skill level associate with soc2010 occupation classifications. See https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2010/soc2010volume1structureanddescriptionsofunitgroups

### 5.4.3 Choosing different training data

The analysis has been repeated a further two times in order to test the sensitivity of the results to the training datasets. As with the main analysis, in all cases the training dataset is composed of 20% of the whole sample, while the testing dataset is composed of the remaining 80% of the sample. In all cases, the sensitivity parameter that determines the complexity of the tree is chosen as the optimal sensitivity parameter for the main tree.[21] This is to allow closer comparability between trees.

Although the precise groups found by the decision tree differ according to the training dataset used, the variables that emerge as most important in determining minimum wage status are similar. In particular, at the root node occupation emerges each time as the first and most important variable in determining workers' minimum wage status.

**Across each of the training datasets, occupation, industry, and a worker's highest qualification level emerge as the most important variables in determining minimum wage status.**

Across each of the decision trees based on different training datasets, many of the nodes are split into subsequent nodes using both the same variable and the same split within that variable as in the main decision tree. In many more cases, the same variable is used to make analogous splits across decision trees, but the precise split chosen within the variable differs slightly. In some cases (especially further down the decision tree, such as at the fourth split and beyond), a different variable is used to split certain nodes. In other cases, some splits further down the decision tree that occur in the main tree do not occur in the alternative trees, while some new splits do emerge.

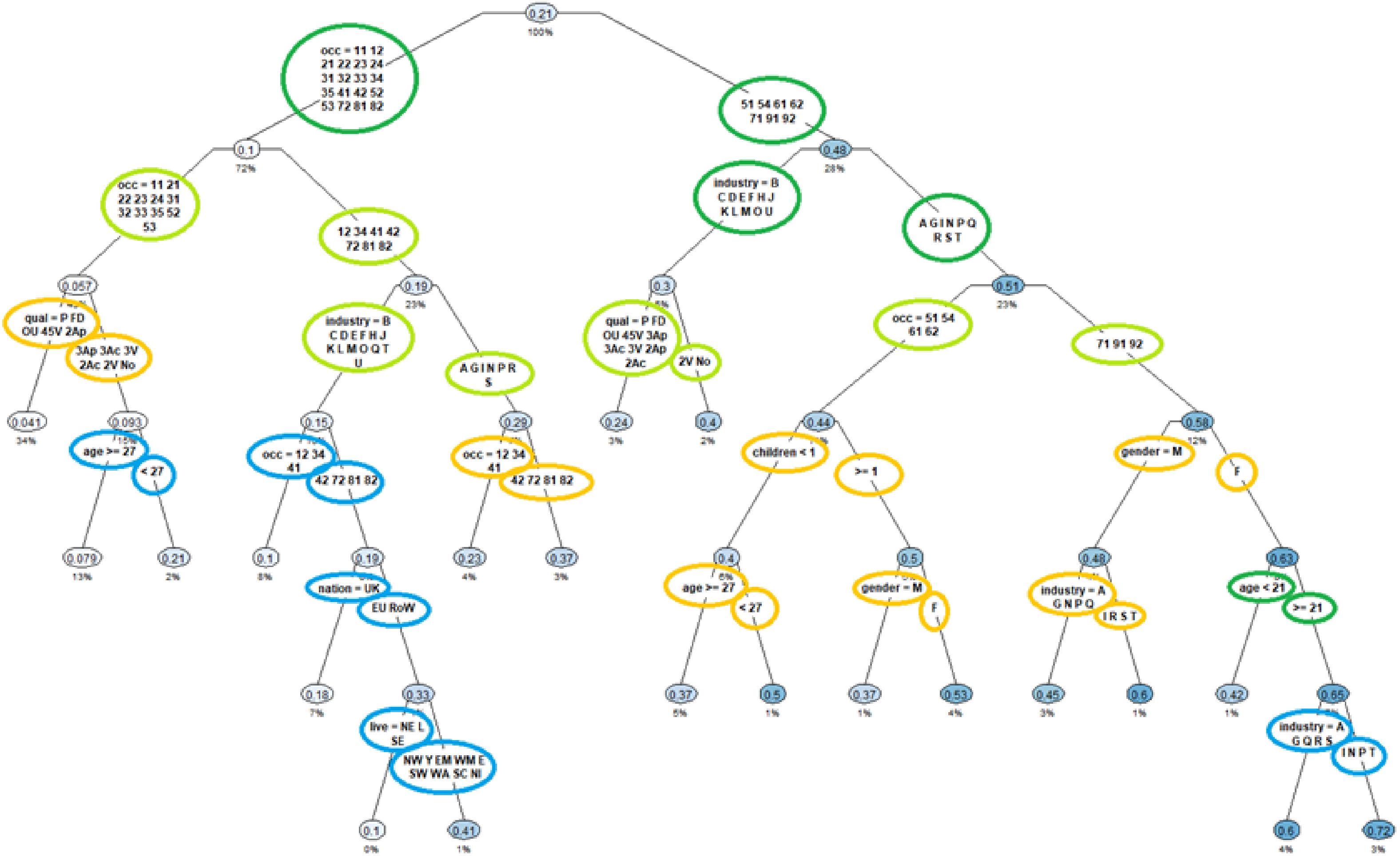
The precise composition of the groups identified by the decision tree is therefore relatively variable depending on the training dataset used. However, although the precise groups may change according to the training dataset used, an ensemble method such as a random forest can be used to validate the probability estimates (and therefore predictions) of the single decision tree model.

Figure 8 and Figure 9 present decision trees produced when using different training data samples from the data. One limitation of decision tree analysis is its sensitivity to different training data, so it is important to understanding which characteristics are consistently most important in predicting whether someone is a minimum wage worker or not.

Across both of these alternative training datasets, early splits within the sample are made using the same variables as those in the decision tree presented in the main report. Occupation emerges as the most important variable in determining minimum wage status, and in both alternative cases is selected by the decision tree to make the first split. Industry again emerges as an important variable in determining minimum wage status.

[21] See Annex **Error! Reference source not found.** for more information.

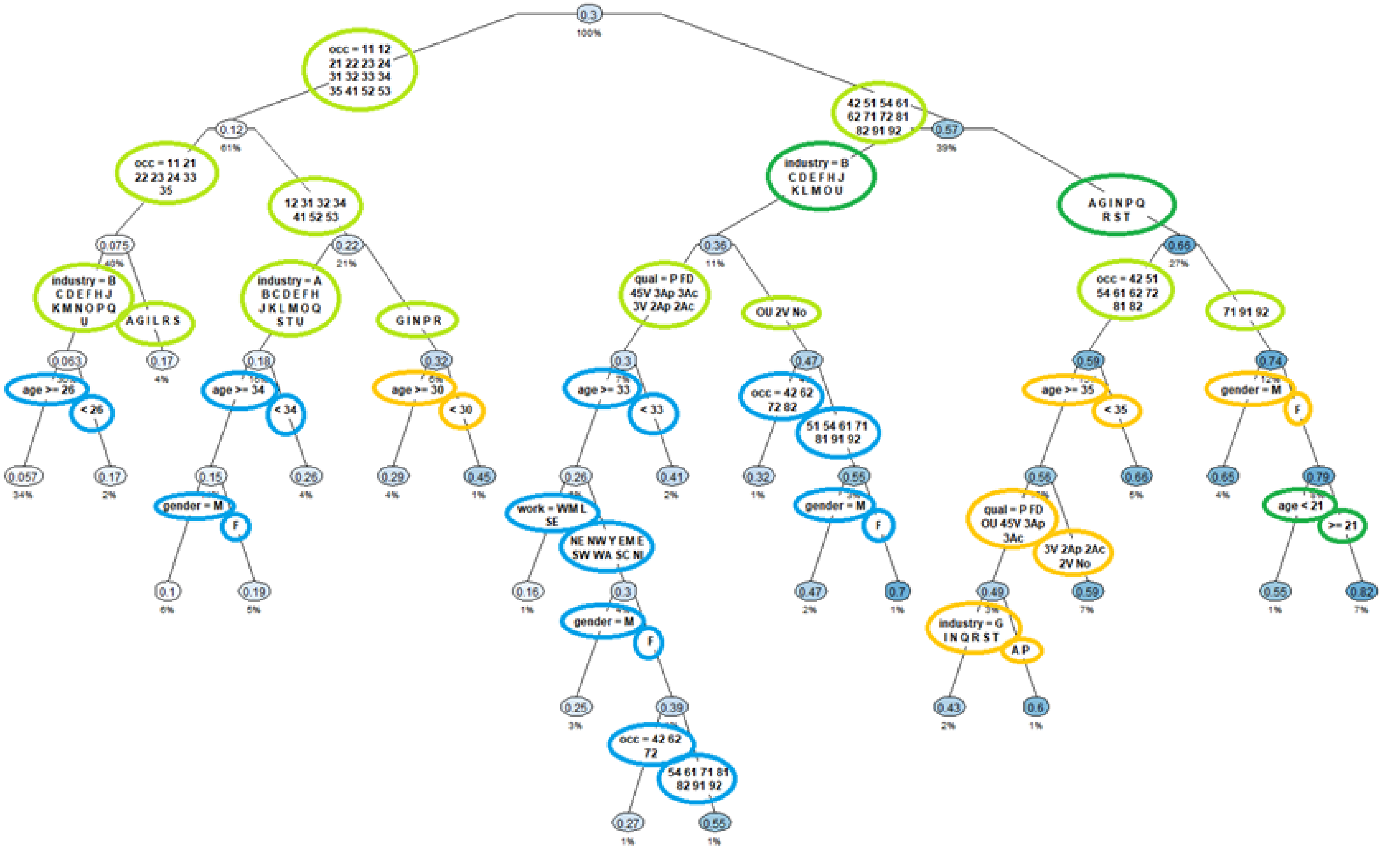
## Figure 5 Decision tree with minimum wage workers defined as those earning 110% of their relevant minimum wage or below



Note: The number inside of the node is the proportion of workers at that node who are minimum wage workers, while the percentage underneath is the proportion of all workers in the sample contained within the node.

***Source: London Economics analysis of LFS data***

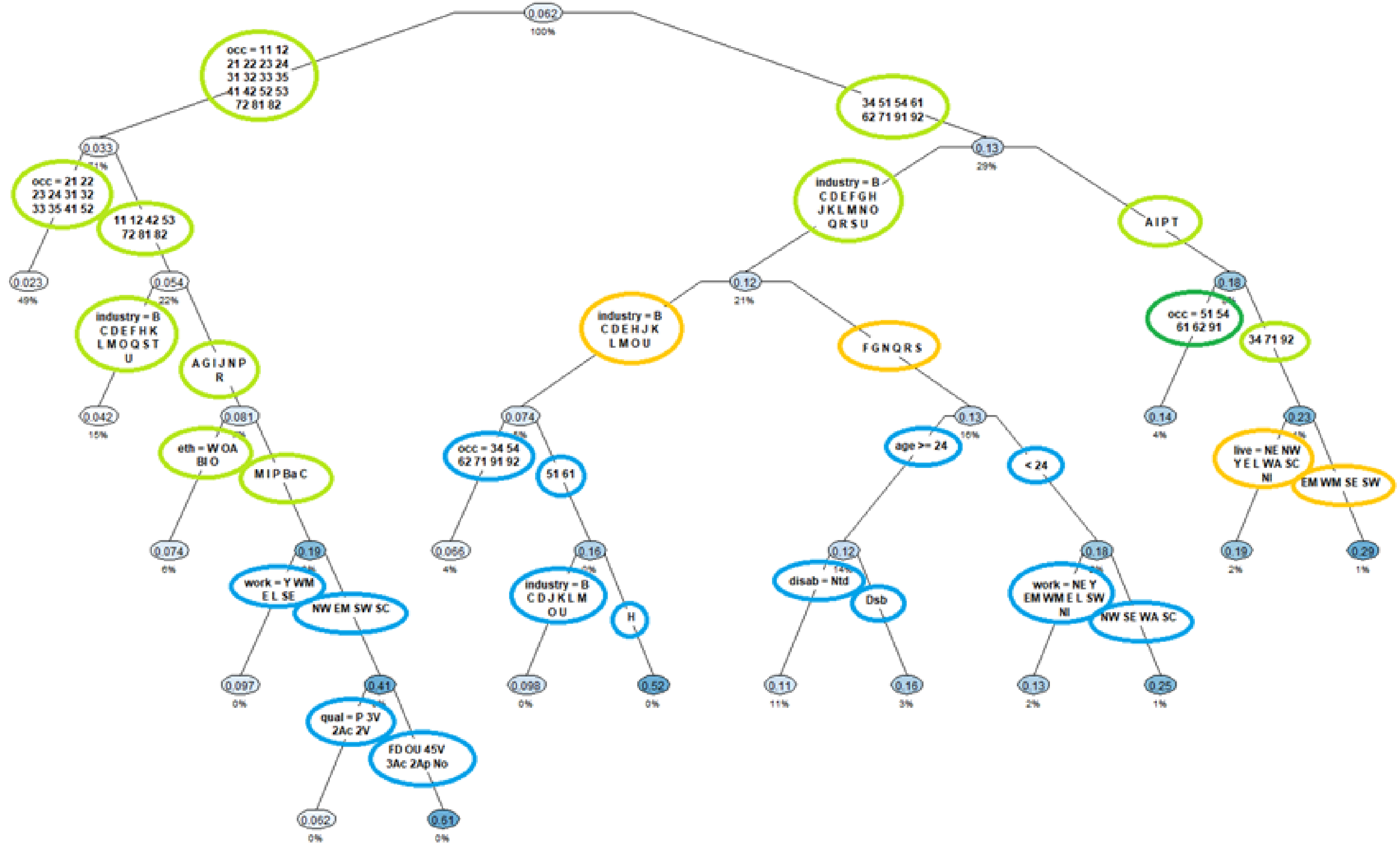
## Figure 6 Decision tree with minimum wage workers defined as those earning 125% of their relevant minimum wage or below



Note: The number inside of the node is the proportion of workers at that node who are minimum wage workers, while the percentage underneath is the proportion of all workers in the sample contained within the node.

***Source: London Economics analysis of LFS data***

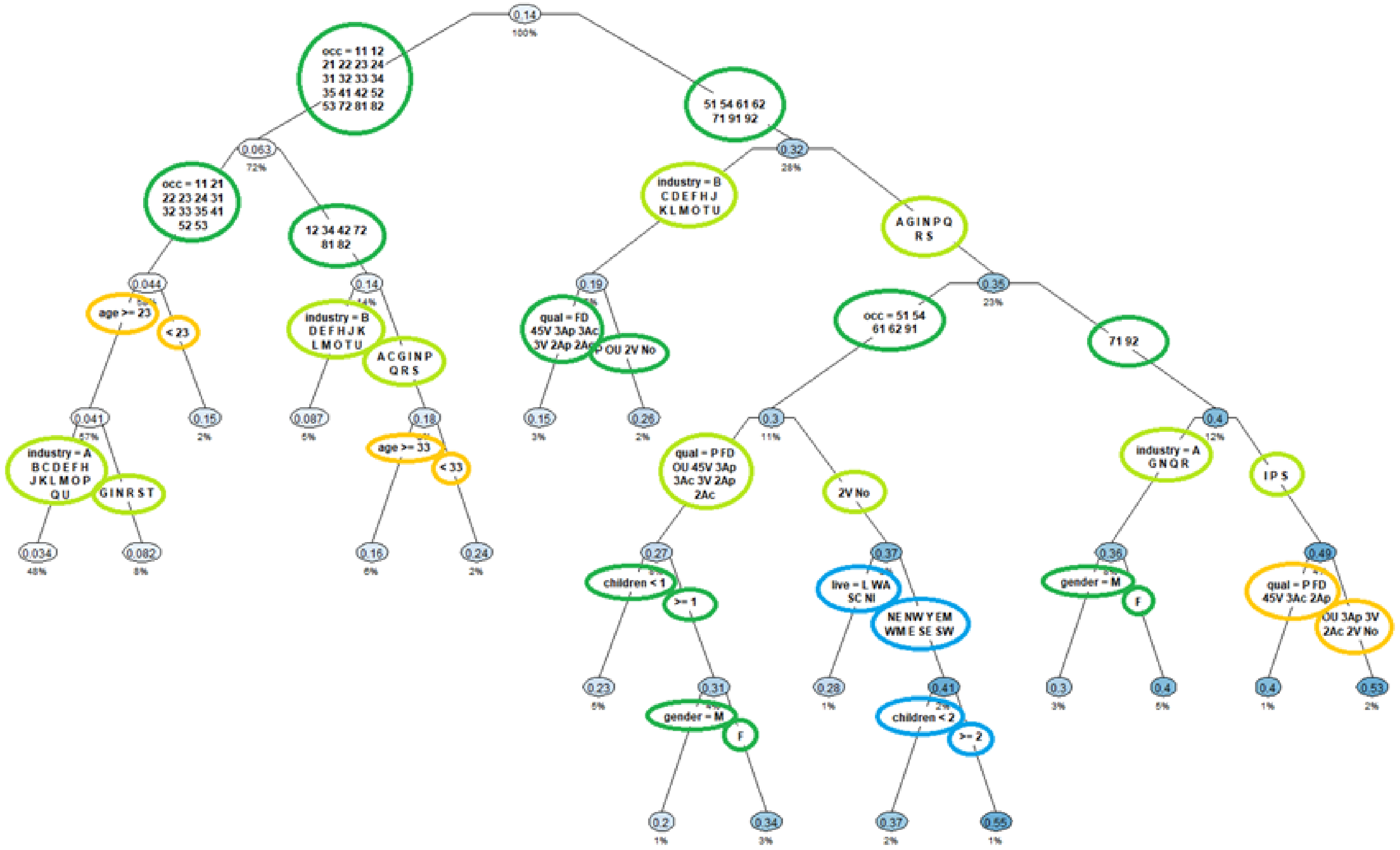
## Figure 7 Decision tree with minimum wage workers defined such that overall coverage is similar to that of the Annual Survey of Hours and Earnings



Note: The number inside of the node is the proportion of workers at that node who are minimum wage workers, while the percentage underneath is the proportion of all workers in the sample contained within the node.

***Source: London Economics analysis of LFS data***

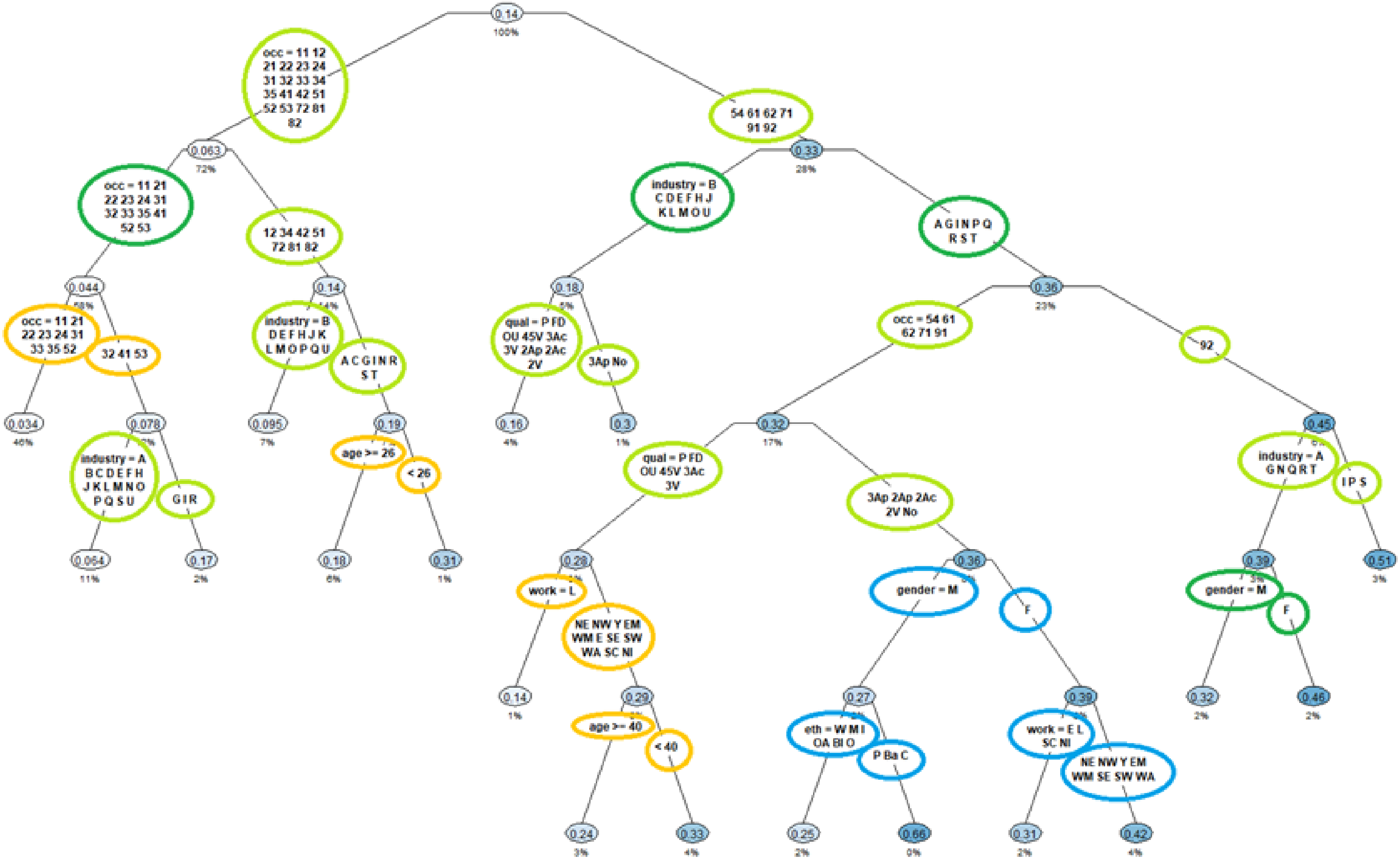
## Figure 8 Decision tree with alternative training data (Example 1)



Note: The number inside of the node is the proportion of workers at that node who are minimum wage workers, while the percentage underneath is the proportion of all workers in the sample contained within the node.

***Source: London Economics analysis of LFS data***

## Figure 9 Decision tree with alternative training data (Example 2)



Note: The number inside of the node is the proportion of workers at that node who are minimum wage workers, while the percentage underneath is the proportion of all workers in the sample contained within the node.

***Source: London Economics analysis of LFS data***

### 5.4.4 Determining the sensitivity parameter used for the decision tree

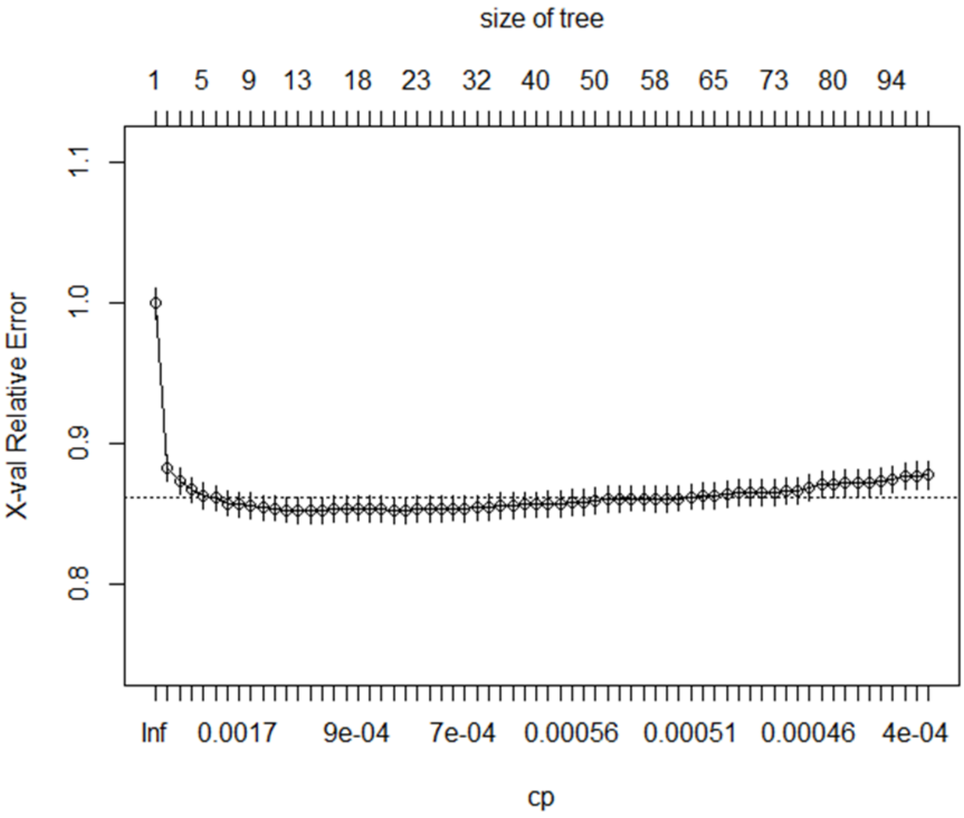
The decision tree analysis has been run for different parameters that impact the number of groups it may identify. While having more groups (leaf nodes) may decrease the error rate, having too many groups also risks overfitting. A loss function that **trades off between the error rate and having many nodes** has been used to determine the "optimal" sensitivity parameter. The optimal sensitivity has been chosen as the sensitivity that minimises the cross-validation relative error.

This analysis identifies the optimal sensitivity parameter (cp) to determine how many splits the decision tree makes. If the decision tree is insufficiently sensitive, it may not undertake important splits in the sample, whereas if it is overly sensitive, then the decision tree may overfit (i.e., create splits that are based on noise in the data that will reduce its ability to predict outcomes).

The cross-validation relative error (X-val Relative Error) measures the average error. This error is typically measured using k-fold cross-validation, where the original data set is split into k subsets. The model is then trained on k-1 subsets and tested on the remaining subset. This process is repeated k times, with each fold used once as the testing set, and the results are averaged to obtain an estimate of the model's prediction error. The optimal sensitivity parameter is the sensitivity value that minimises the cross-validation relative error.

In Figure 10Figure 10, having no splits in the decision tree (only one node) is normalised to one.

**Figure 10 Finding the optimal sensitivity of the decision tree**



Note: The cross-validation relative error (X-val Relative Error) measures the average error (normalised as one when there is only one node), while 'cp' is the sensitivity of the decision tree algorithm. The greater the sensitivity the more splits will be made, so determining the sensitivity is equivalent to determining the size of the tree.

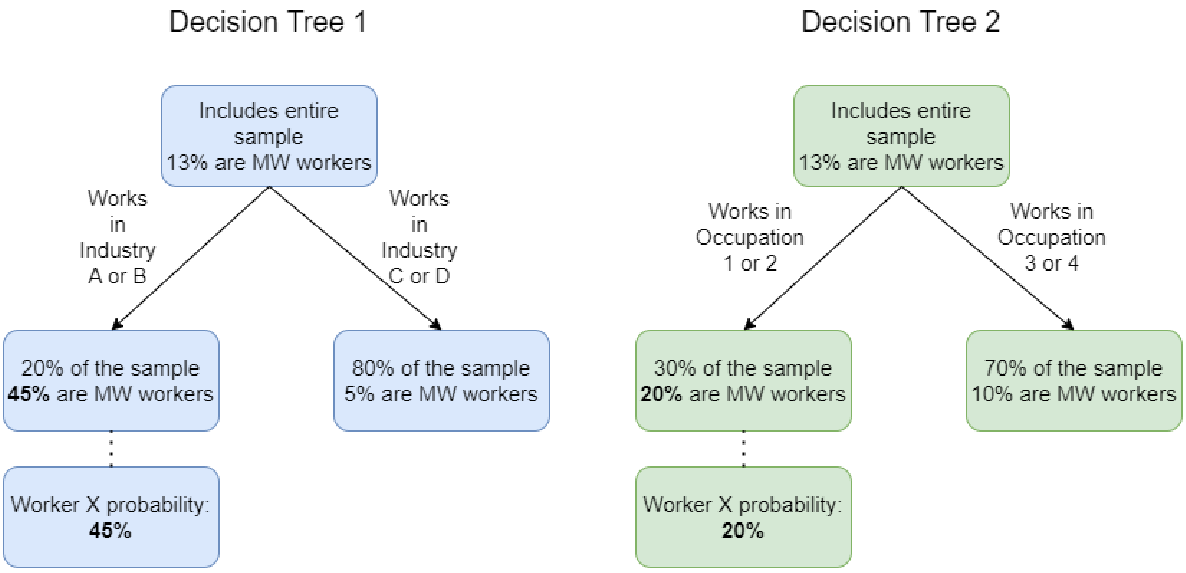
***Source: London Economics analysis of LFS data***

### 5.4.5 Ensemble methods: random forests and boosted trees

One criticism of decision trees is that they may be sensitive to which portion of the data is used as the training data. Characteristics that are deemed critical in splitting workers apart in one training sample may not be in another. Although robustness checks from the previous section suggests limited sensitivity to changing the training sample, ensemble methods (random forests and boosted gradient trees) are used to confirm this. Ensemble methods such as random forests and boosted gradient trees use multiple decision trees. The estimated probabilities of workers being a minimum wage worker is estimated using an average of the probabilities estimated by each decision tree, which are separately estimated.

The following is an example of a **simple random forest** made up two simple decision trees. Each decision tree is estimated using a different subsample of the data and provide a single split of the sample into two groups.

**Figure 11 Example simple random forest with two decision trees**


Decision Tree 1
Decision Tree 2
Includes entire sample 13% are MW workers
Works in Industry A or B
Works in Industry C or D
20% of the sample **45%** are MW workers
80% of the sample 5% are MW workers
Worker X probability: **45%**
Includes entire sample 13% are MW workers
Works in Occupation 1 or 2
Works in Occupation 3 or 4
30% of the sample **20%** are MW workers
70% of the sample 10% are MW workers
Worker X probability: **20%**

Note: The above diagram is solely for illustrative purposes and does not reflect the data from the LFS.

Figure 11 presents how a random forest estimates the probability that each worker is a minimum wage worker. The simple random forest (solely for illustrative purposes and does not reflect real-world data) contains two decision trees, which use different training samples from the data. The first (Decision Tree 1) prioritises splitting workers into groups based on the industry they work in, whereas the second (Decision Tree 2) prioritises splitting workers into groups based on the occupation they work in.

A worker X who works in Industry A an Occupation 1 will be assigned to the left leaf node (group) in Decision Tree 1, where 45% of workers are minimum wage workers. As a result, the estimated probability that worker X is a minimum wage worker according to Decision Tree 1 is 45%. Similarly, Decision Tree 2 estimates that worker X is a minimum wage worker with probability 20%. The random forest of two decision trees takes the average of the two probabilities, resulting in the random forest estimating a probability of 32.5%.

An alternative method to predict minimum wage status is **gradient boosted trees**. This is also an ensemble method.

This is the preferred method of Cengiz et al. (2022), as it

1) provides a more robust estimate than a single decision tree, and
2) provides a potentially more precise estimate than random forests, as the ensemble of trees are built sequentially – this allows for the algorithm to improve its predictions.

The creation of the nth decision tree in a sequence of trees (i.e., gradient boosted trees) is dependent on previous n-1 trees. The nth decision tree focuses on correcting the previous n-1 trees. As described previously, decision trees minimise a loss function (entropy).

Among gradient boosted trees, the loss function that the nth decision tree minimises places greater weighting on workers that have been most misclassified in the previous n-1 trees. For example, a worker that is a minimum wage worker and has been assigned to a group where only 1% of workers are minimum wage workers is more misclassified than a minimum wage worker that has been assigned to a group where 50% of workers are minimum wage workers (and vice versa for non-minimum wage workers).

**Results**

Although it is **not feasible to show all the decision trees within the random forest or gradient boosted trees**, we can estimate the **relative importance of different variables**, following the methodology (Friedman, 2001) that Cengiz et al. (2022) use. This produced in Figure 12 (random forest) and Figure 13 (gradient boosted trees).

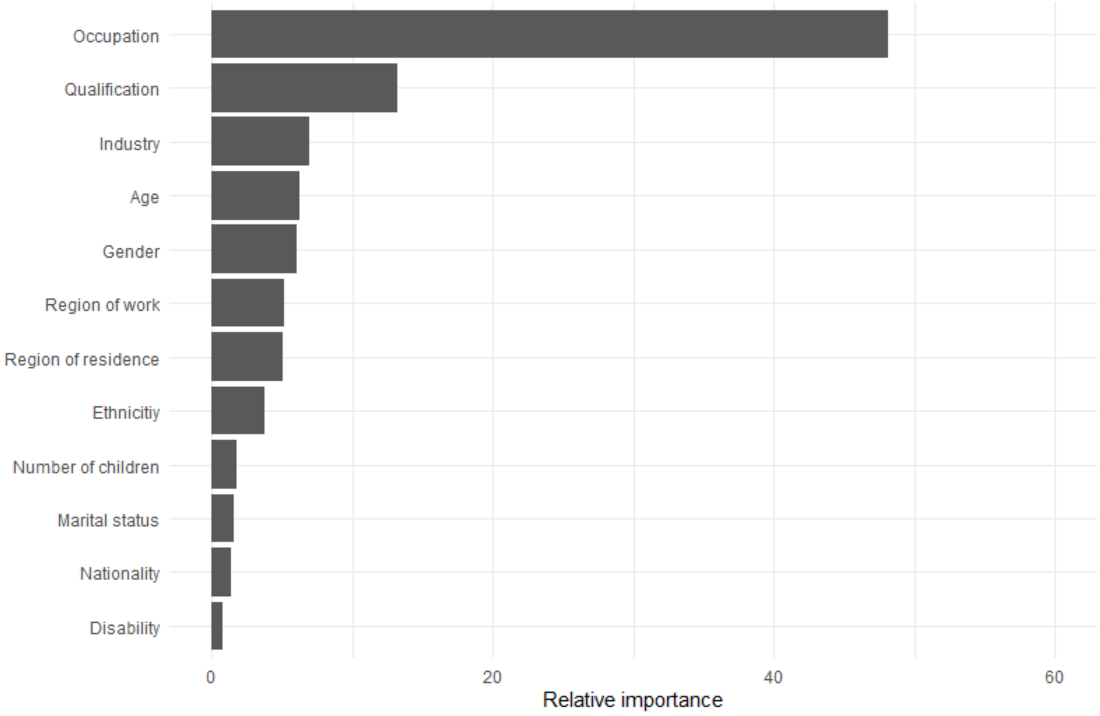
The bars represent the decline in the loss function associated with the variable if it were excluded, with the bars normalised to sum across variables to 100. These estimates can be interpreted as **how much the model improves** with the inclusion of a particular variable.

Both the random forest and gradient boosted trees suggest that occupation is by the most important characteristic when explaining whether a worker is a minimum wage worker. For the random forest the highest qualification of the worker is the second most important characteristic followed their industry of work, while it is the other way around for gradient boosted trees.

Gradient boosted trees place a greater emphasis on occupation than random forest, illustrated by a large relative variable importance. However, the order of relative importance of variables is largely similar. For example, nationality and disability are estimated to be less important in predicting whether a worker is a minimum wage worker.

This is consistent with the decision tree results as presented. Occupation is always the first split taken by the decision tree, followed by industry and qualification. Other variables, such as gender and number of children appear lower down in these decision trees, if at all.

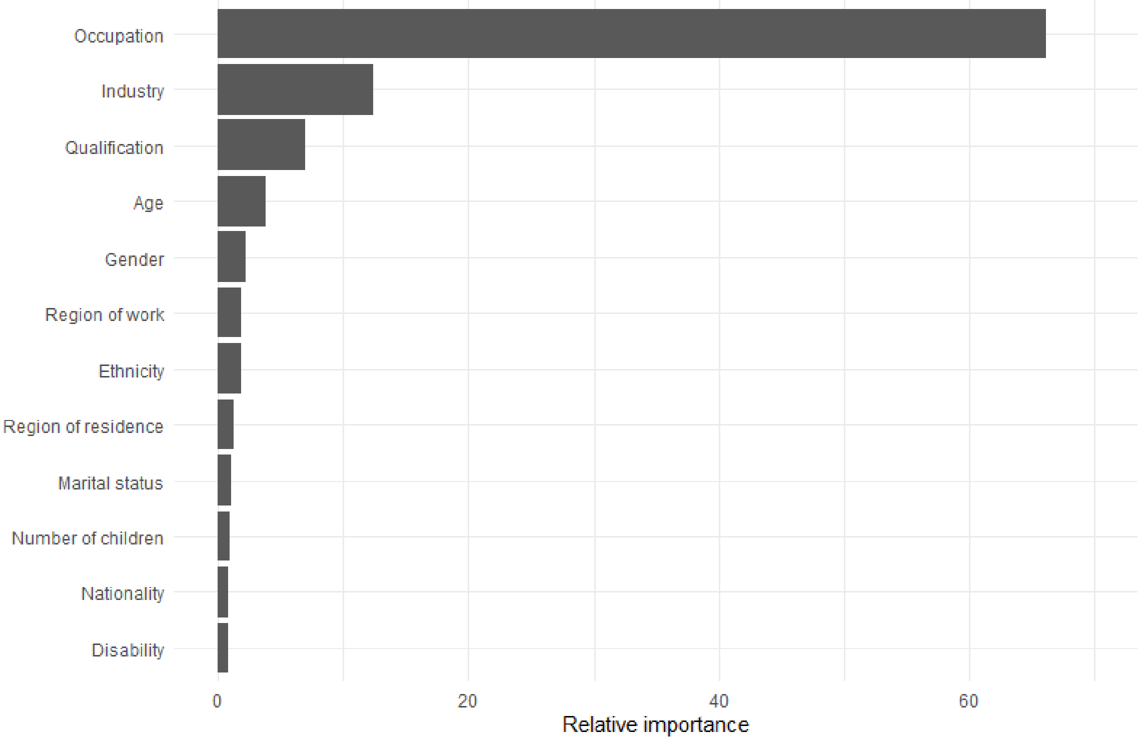
## Figure 12 Relative variable importance – random forest



Note: Relative importance is normalised such that the sum of the relative importance across characteristics is 100.

***Source: London Economics analysis of LFS data***

## Figure 13 Relative variable importance – gradient boosted trees



Note: Relative importance is normalised such that the sum of the relative importance across characteristics is 100.

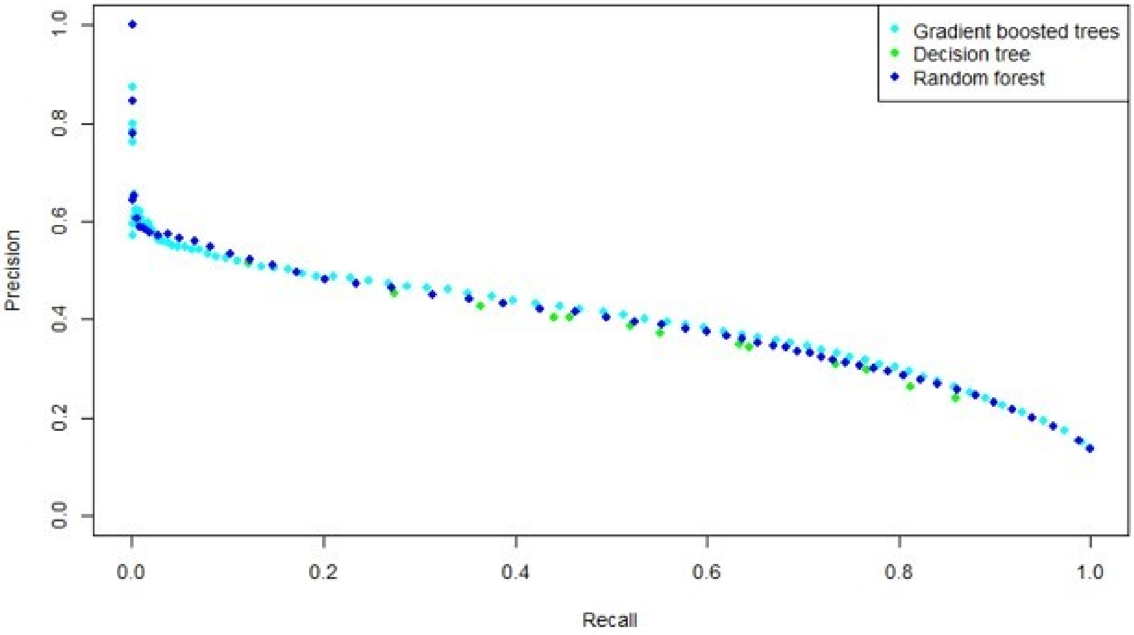
***Source: London Economics analysis of LFS data***

## Evaluation

Figure 14 and Figure 15 present the precision and recall of the gradient boosted trees and random forest compared to the decision tree. Figure 14 presents the full precision-recall curve, while Figure 15 focuses on the part of the curve with precision lower than 50% (higher recall) to illustrate the differences in the performance of the three tree-based classifications.

As expected, the random forest (in dark blue) outperforms the decision tree (green): the random forest is able to identify a given proportion of minimum wage workers (a given recall) more precisely. Gradient boosted trees (light blue) outperforms both random forest and decision tree.

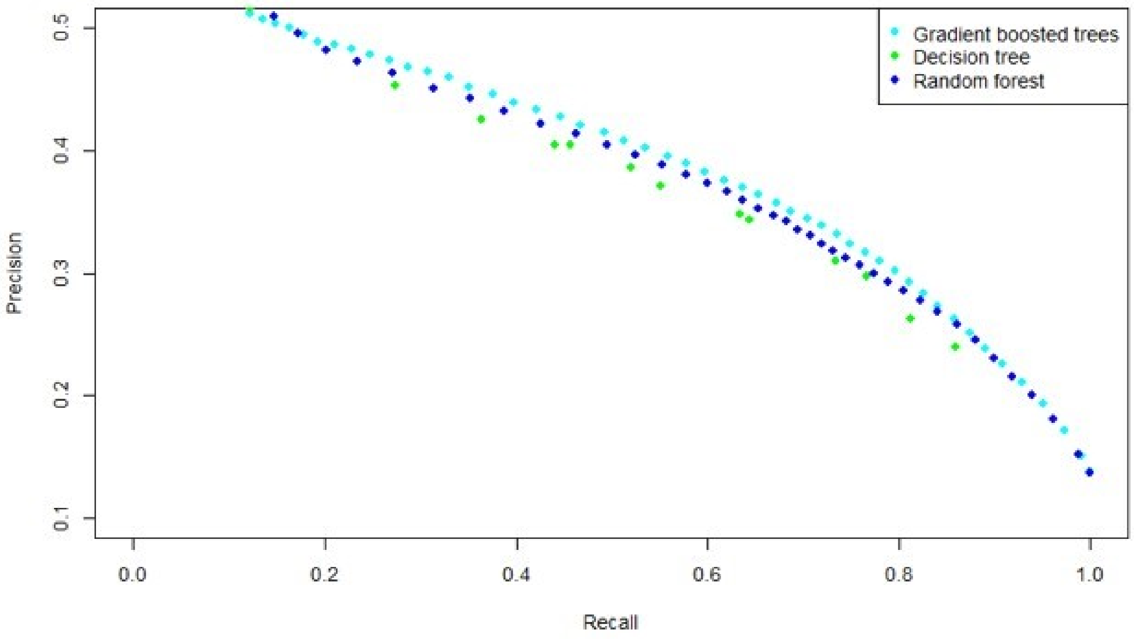
### Figure 14 Precision-recall comparison



Note: Precision indicates the probability that a randomly chosen member of a group is a minimum wage worker. Recall indicates the proportion of all minimum wage workers that are contained within that group. Each numerical label specifies a probability threshold that is used to determine whether a worker is predicted to be a minimum wage worker or not.

***Source: London Economics analysis of LFS data***

**Figure 15 Precision-recall comparison focusing on precision lower than 50%**

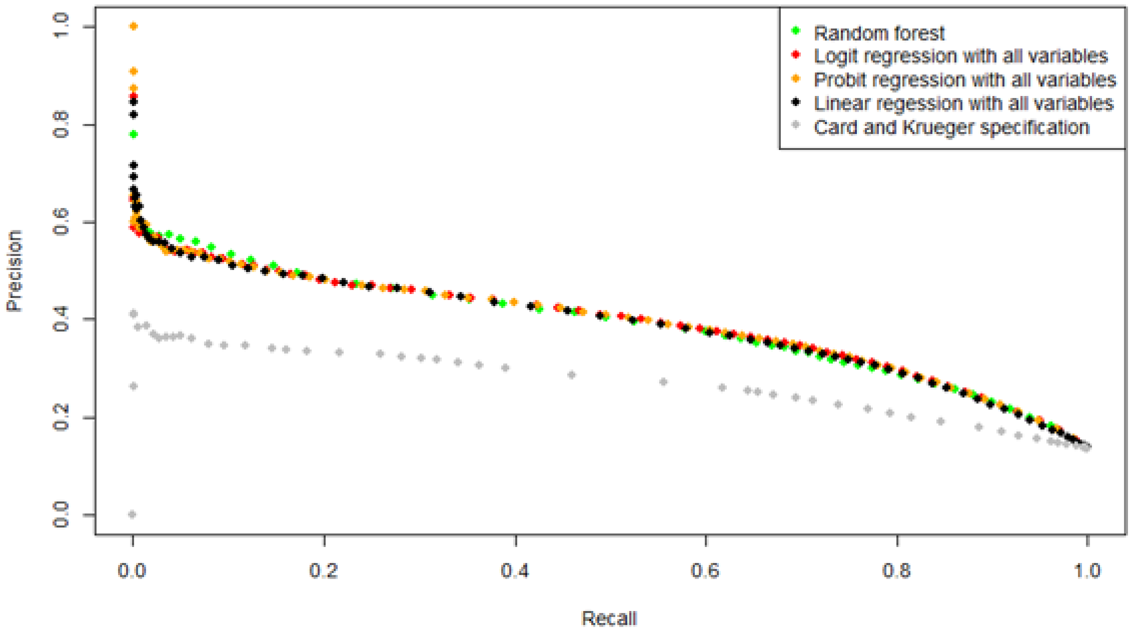


Note: Precision indicates the probability that a randomly chosen member of a group is a minimum wage worker. Recall indicates the proportion of all minimum wage workers that are contained within that group. Each numerical label specifies a probability threshold that is used to determine whether a worker is predicted to be a minimum wage worker or not.

***Source: London Economics analysis of LFS data***

In addition, these methods can be compared to other prediction models besides tree-based classification methods, as presented in Section 5.3. Figure 16 and Figure 17 compare precision-recall curves of random forest and gradient boosted trees, respectively, and other regression predictors. Performing better than decision trees, both methods perform similarly well as the logit/probit/linear models and considerably better than the Card and Krueger specification.

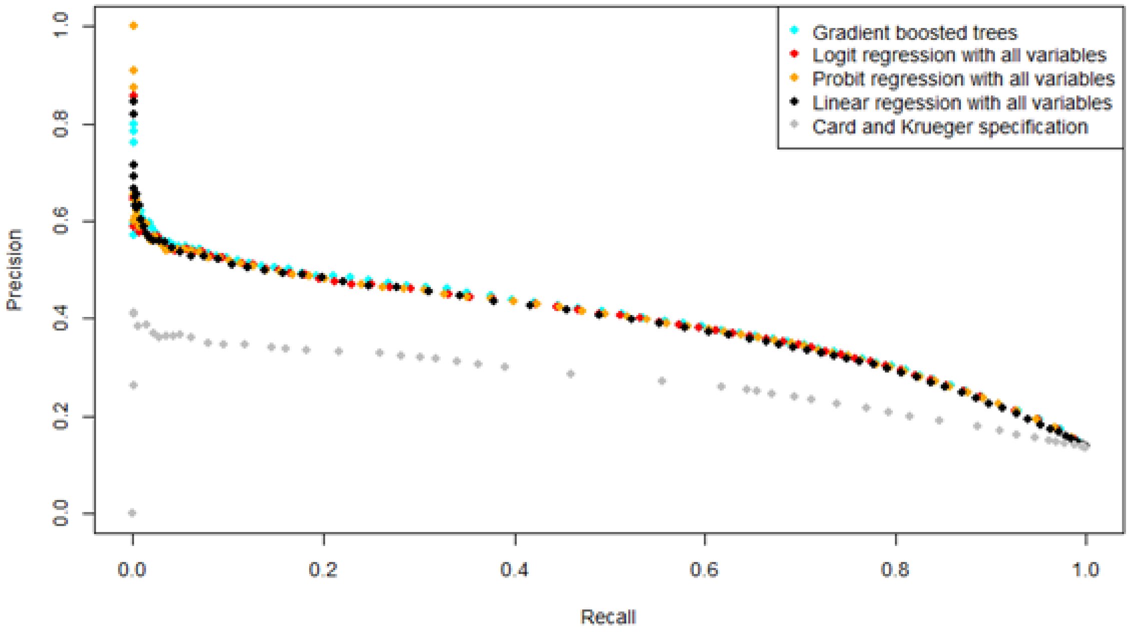
## Figure 16 Precision-recall comparison between random forest and regression predictors



Note: Precision indicates the probability that a randomly chosen member of a group is a minimum wage worker. Recall indicates the proportion of all minimum wage workers that are contained within that group. Each numerical label specifies a probability threshold that is used to determine whether a worker is predicted to be a minimum wage worker or not.

***Source: London Economics analysis of LFS data***

## Figure 17 Precision-recall comparison between gradient boosted trees and regression predictors



Note: Precision indicates the probability that a randomly chosen member of a group is a minimum wage worker. Recall indicates the proportion of all minimum wage workers that are contained within that group. Each numerical label specifies a probability threshold that is used to determine whether a worker is predicted to be a minimum wage worker or not.

***Source: London Economics analysis of LFS data***

# 6 Identifying potential minimum wage workers

## 6.1 Methodology

The above analysis focused on workers who are **employees** - i.e., those where information about job characteristics (occupation and sector) are available. It is also important to understand **which unemployed/inactive workers are impacted by changes** to the minimum wage, as they can provide a **counterfactual for those in work who are minimum wage workers.**

**Removing minimum wage workers**

A predictive model without job characteristics can be used to predict the likelihood of someone who is economically inactive/unemployed being a minimum wage worker if they were working. The following analysis removes occupation and sector from the tree-based classification. The index of multiple deprivation (IMD) of the local area (LSOA)[22], migrant status (whether the worker is a migrant or not based on their country of birth), and the rural/urban classification of the local area (as defined by the ONS) are also included in this analysis.

However, it is important to note that although the predictive model could provide predictions as to whether a worker who is unemployed would be a minimum wage worker (i.e., shares characteristics which those who are more likely to be minimum wage workers or not), **it is not possible to test this given that those not in work do not have hourly pay information**.

**Limitations**

Predicting 'potential' minimum wage status of those out of work using information from those in work **assumes that those out of work** (controlling for the characteristics included in the predictive model) **would be similar to those in work in their probability of being a minimum wage worker**.

However, there may be reasons why a worker is unemployed or inactive that is not captured in the characteristics in the predictive model that may influence the probability that the worker would be a minimum wage worker if they were in work. For example, experience (in work) may be associated with both unemployment (as it is more difficult to find work with less prior experience) and whether a worker is a minimum wage worker (with less prior experience a worker may have less bargaining power when negotiating pay).

If those who are unemployed are more likely to have less work experience, then a predictive model that does not include work experience may underestimate the proportion of those who are unemployed who would be minimum wage workers.

Further work would explore the use of information about the previous job that an unemployed or inactive worker was in, and the extent to which conclusions made based on employees can be extrapolated to those out of work.

[22] The IMD variable used in the analysis indicates which quintile their local area is in based on their IMD index.

## 6.2 Results

Similar to Table 1 which outlined the groups identified by the decision tree when job characteristics are included, Table 3 presents the groups identified by the decision tree when job characteristics are not included.

Similar to Table 1, the decision tree categorises a large proportion of the sample into a few large groups that have relatively low concentration of minimum wage workers: around half of the sample are included in the two groups that have the lowest concentration of minimum wage workers. Six out of the seven groups with the highest concentration of minimum wage workers (making up around a fifth of all minimum wage workers) are made up of female workers (the exception being male migrant workers from Pakistani, Bangladeshi, or mixed backgrounds aged 26 and older with relatively low educational qualifications). Educational qualifications appear to be a critical characteristic in predicting whether a worker is a minimum wage worker in the absence of job characteristics, defining all nineteen groups.

### Table 3 Decision tree groups without job characteristics (ranked by precision)

| Rank (by precision) | Precision | Recall | % of sample | Description |
|---|---|---|---|---|
| 1 | 60.6% | 1.2% | 0.3% | Female workers aged 40 and older with Level 2 vocational/no qualifications, living in Yorkshire in a local area in the top quintile of the Index of Multiple Deprivation (most deprived). |
| 2 | 51.8% | 0.9% | 0.3% | Female workers aged 39 and younger with Level 2 vocational/no qualifications with no dependent children, living in Yorkshire or the North East. |
| 3 | 48.0% | 1.1% | 0.3% | Male migrant workers from Pakistani, Bangladeshi, or mixed ethnic backgrounds aged 26 and older with a Level 2/Level 3/other degree/no qualifications. |
| 4 | 44.8% | 5.8% | 2.0% | Female workers aged 39 and younger with Level 2 vocational/no qualifications with dependent children. |
| 5 | 39.2% | 4.4% | 1.7% | Female workers with other degree/Level 3 academic or vocational/Level 2 academic qualifications with dependent children working in the North East, East of England, South East, South West, Wales, and Scotland living in a local area in the top two quintiles of the Index of Multiple Deprivation (two most deprived). |
| 6 | 33.6% | 2.6% | 1.2% | Female workers aged 40 and older with Level 2 vocational/no qualifications, living in regions other than Yorkshire in a local area in the top quintile of the Index of Multiple Deprivation (most deprived). |
| 7 | 29.3% | 4.5% | 2.3% | Female workers aged 27 and younger with other degree/Level 3 academic or vocational/Level 2 academic qualifications living in a local area in the bottom three quintiles of the Index of Multiple Deprivation (three least deprived). |
| 8 | 28.6% | 7.0% | 3.7% | Male workers aged 25 or younger with other degree/Level 3 academic or vocational/Level 2 academic or vocational/no qualifications. |
| 9 | 27.2% | 6.8% | 3.8% | Female workers aged 40 and older with Level 2 vocational/no qualifications living in a local area in the top quintile of the Index of Multiple Deprivation (most deprived). |

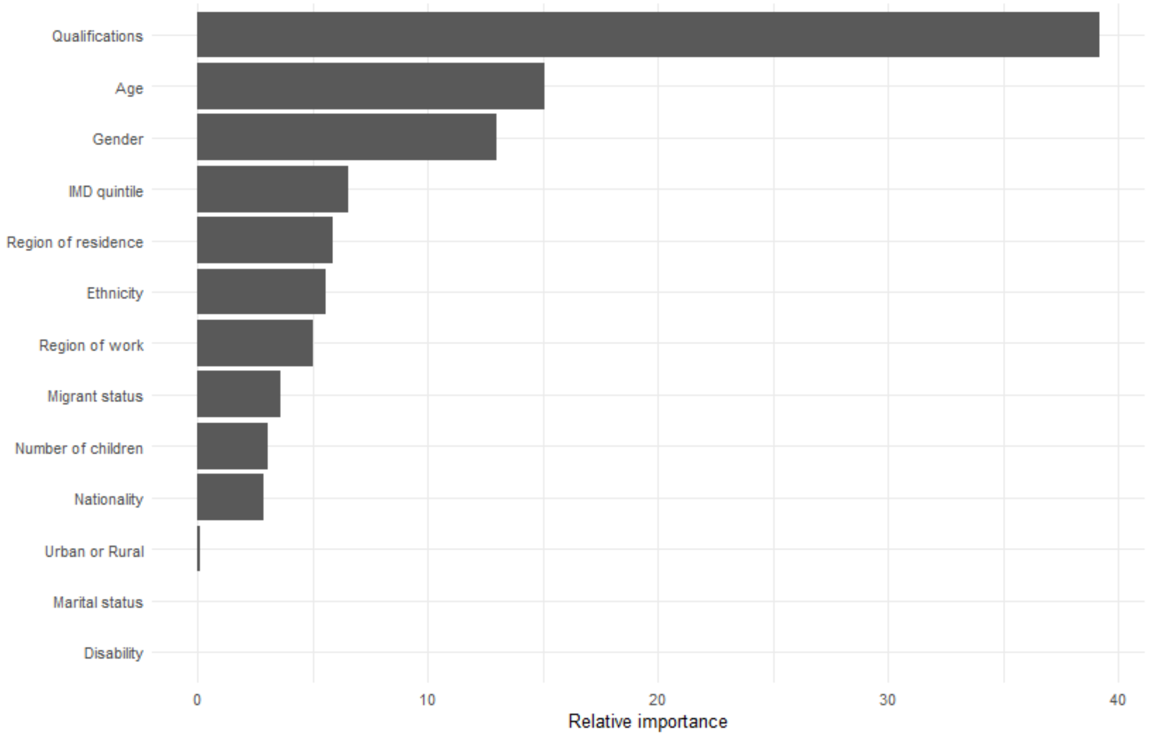
| Rank (by precision) | Precision | Recall | % of sample | Description |
|---|---|---|---|---|
| 10 | 25.6% | 3.2% | 1.9% | Female workers with other degree/Level 3 academic or vocational/Level 2 academic qualifications with dependent children working in the North West, Yorkshire, East Midlands, West Midlands, London, and Northern Ireland living in a local area in the top two quintiles of the Index of Multiple Deprivation (two most deprived). |
| 11 | 24.1% | 8.1% | 5.1% | Female workers with other degree/Level 3 academic or vocational/Level 2 academic qualifications with no dependent children living in a local area in the top two quintiles of the Index of Multiple Deprivation (two most deprived). |
| 12 | 22.1% | 1.1% | 0.8% | Female workers aged 39 and younger with Level 2 vocational/no qualifications with no dependent children, living in regions other than Yorkshire or the North East. |
| 13 | 20.6% | 2.4% | 1.8% | Workers aged 23 and younger with postgraduate/first degree/Level 4 or 5 vocational/Level 2 or 3 apprenticeship qualifications. |
| 14 | 19.4% | 4.0% | 3.1% | Male migrant workers from White, Indian, Chinese, other Asian, Black/African/Caribbean/Black British, or other ethnic backgrounds aged 26 and older with other degree/Level 3 academic or vocational/Level 2 academic or vocational/no qualifications. |
| 15 | 19.1% | 1.5% | 1.2% | Workers from Bangladeshi, other Asian, or mixed ethnic backgrounds aged 24 and above with postgraduate/first degree/Level 4 or 5 vocational/Level 2 or 3 apprenticeship qualifications. |
| 16 | 16.5% | 5.6% | 5.1% | Male non-migrant workers aged 26 or older with Level 2 vocational/no qualifications. |
| 17 | 15.3% | 11.9% | 11.7% | Female workers aged 28 and older with other degree/Level 3 academic or vocational/Level 2 academic qualifications living in a local area in the bottom three quintiles of the Index of Multiple Deprivation (three least deprived). |
| 18 | 8.4% | 8.1% | 14.5% | Male non-migrant workers aged 26 or older with other degree/Level 3 academic or vocational/Level 2 academic qualifications. |
| 19 | 7.5% | 19.6% | 39.3% | Workers from White, Indian, Pakistani, Chinese, Black/African/Caribbean/Black British, other ethnic backgrounds O aged 24 and above with postgraduate/first degree/Level 4 or 5 vocational/Level 2 or 3 apprenticeship qualifications. |

Note: Precision indicates the probability that a randomly chosen member of a group is a minimum wage worker. Recall indicates the proportion of all minimum wage workers that are contained within that group. The recall and percentage of sample figures presented are rounded to the nearest 0.1%, so the sum of the rounded estimates may not equal 100.0%. Qualifications refer to the highest-level qualification attained.

***Source: London Economics analysis of LFS data***

Figure 18 presents the relative variable importance of the decision tree without job characteristics. Consistent with the decision tree analysis that did include job characteristics presented in Section 5.2, highest-level qualification is the most important characteristic, followed by age and gender. The ordering of the variables by relative variable importance is consistent with the ordering presented in Figure 12. For example, disability and marital status remain relatively less important in predicting whether a worker is a minimum wage worker. The deprivation in the local area within which a worker lives is relatively important (potentially as prices and wages in the local area may be lower in more deprived areas), although the other additional characteristics (migrant status and rural/urban classification) are less important.

**Figure 18 Relative variable importance without job characteristics – decision tree**



Note: Relative importance is normalised such that the sum of the relative importance across characteristics is 100.

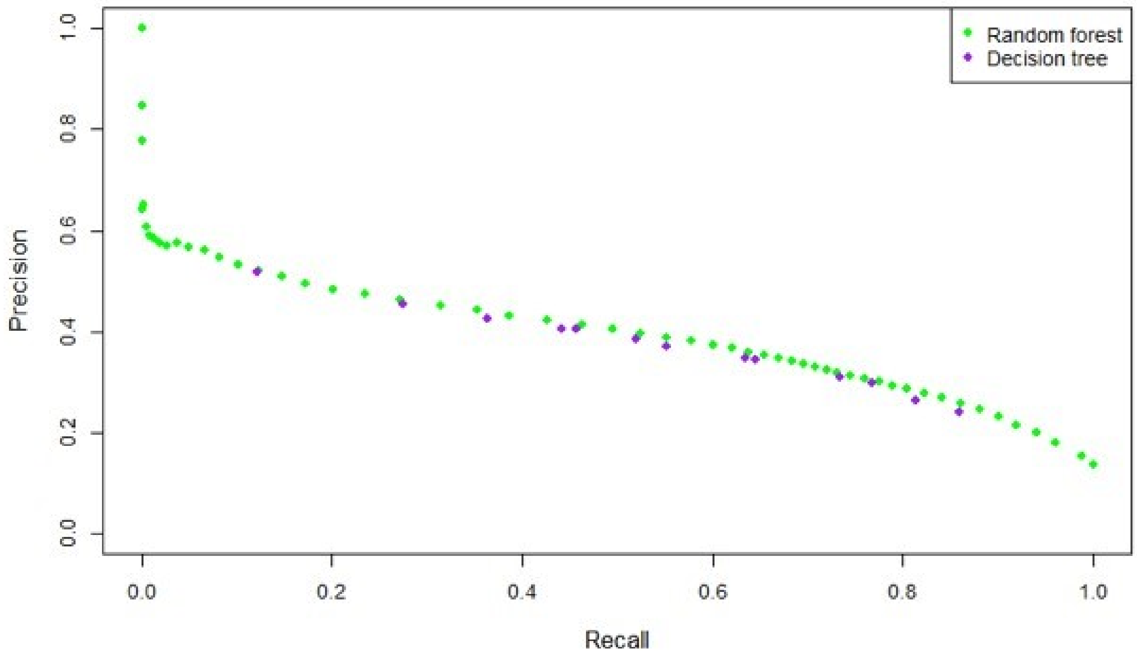
***Source: London Economics analysis of LFS data***

## 6.3 Evaluation

Figure 19 and Figure 20 present the precision-recall curves of a decision tree and random forest (as a robustness check of the evaluation). Figure 19 presents the precision-recall curves of a decision tree/random forest that include job characteristics, while Figure 20 presents the precision-recall curves of a decision tree/random forest that do not include job characteristics. The comparison between Figure 19 and Figure 20 suggests that the exclusion of job characteristics reduces the predictive power of both the decision tree and random forest. For a given recall, including job characteristics allows for a higher precision to be achieved.

This is consistent with occupation's relative variable importance as presented in Figure 12. Occupation's relative variable importance when included is far greater than other characteristics while it is the first characteristic that the decision tree splits the sample by across a range of specifications, so the decrease in predictive power of the decision tree/random forest in response to the exclusion of occupation from the model is expected.

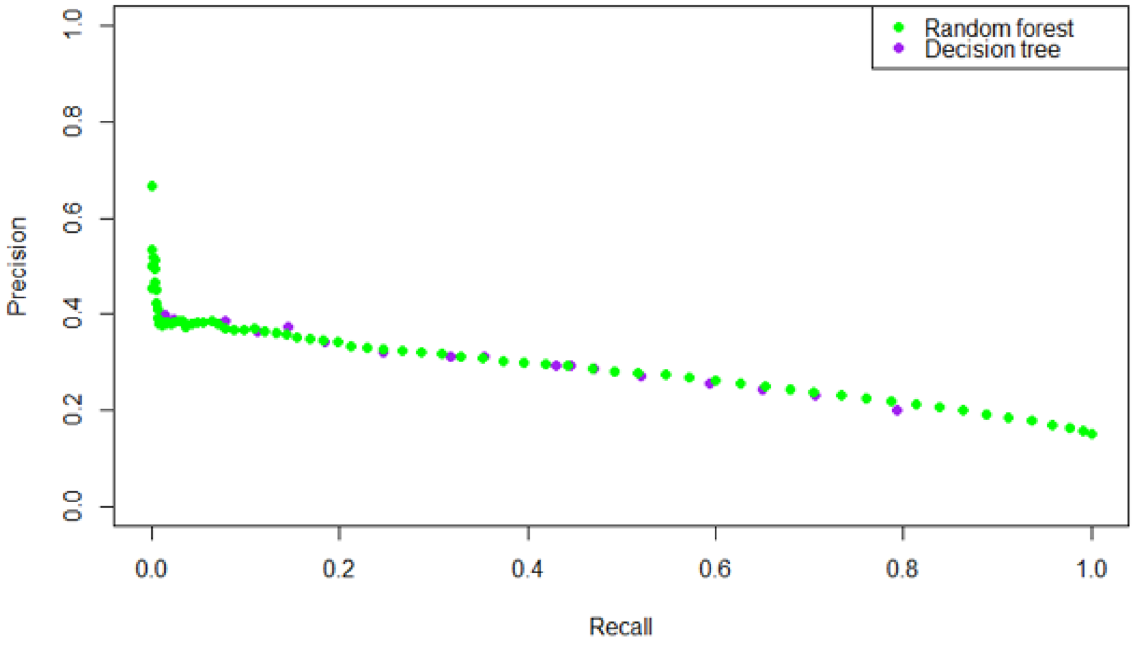
## Figure 19 Precision-recall curves including job characteristics



Note: Precision indicates the probability that a randomly chosen member of a group is a minimum wage worker. Recall indicates the proportion of all minimum wage workers that are contained within that group. Each numerical label specifies a probability threshold that is used to determine whether a worker is predicted to be a minimum wage worker or not.

***Source: London Economics analysis of LFS data***

## Figure 20 Precision-recall excluding job characteristics



Note: Precision indicates the probability that a randomly chosen member of a group is a minimum wage worker. Recall indicates the proportion of all minimum wage workers that are contained within that group. Each numerical label specifies a probability threshold that is used to determine whether a worker is predicted to be a minimum wage worker or not.

***Source: London Economics analysis of LFS data***

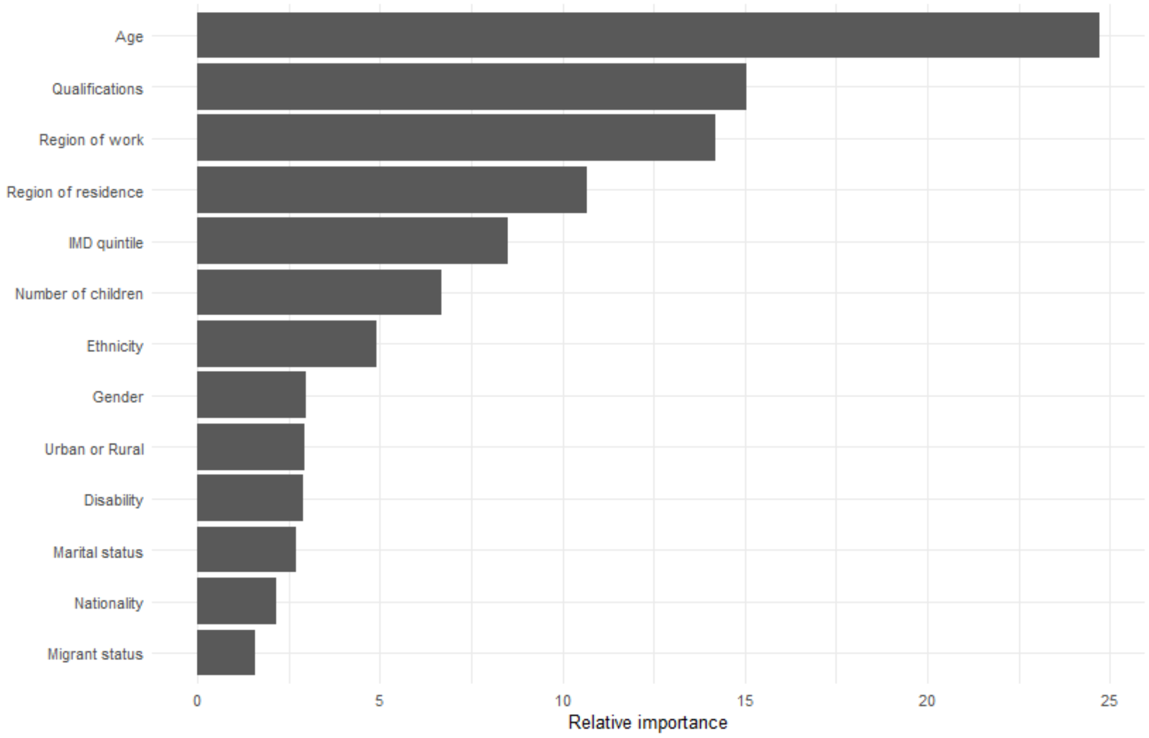
## 6.4 Sensitivity and robustness checks

As suggested in the previous section, a random forest is used to test the robustness of the conclusions drawn from the decision tree. Figure 21 presents the relative variable importance of a random forest when job characteristics are not included.

Consistent with the decision tree findings, age and qualifications are the most important predictors (although their ordering switches), while region of residence and the local area's level of deprivation are also relatively important in both the decision tree and the random forest.

One key difference between the random forest and the decision tree analysis is that the relative importance is more evenly spread across characteristics in the random forest. This suggests that qualifications and age were particularly important in the particular sample from which the decision tree was trained on. This difference highlights the importance of drawing conclusions from a range of decision trees, although it reduces the interpretability of its findings (such as identifying interaction effects between characteristics).

**Figure 21 Relative variable importance – random forest**



Note: Relative importance is normalised such that the sum of the relative importance across characteristics is 100.

***Source: London Economics analysis of LFS data***

# 7 Latent Dirichlet Allocation

## 7.1 Methodology

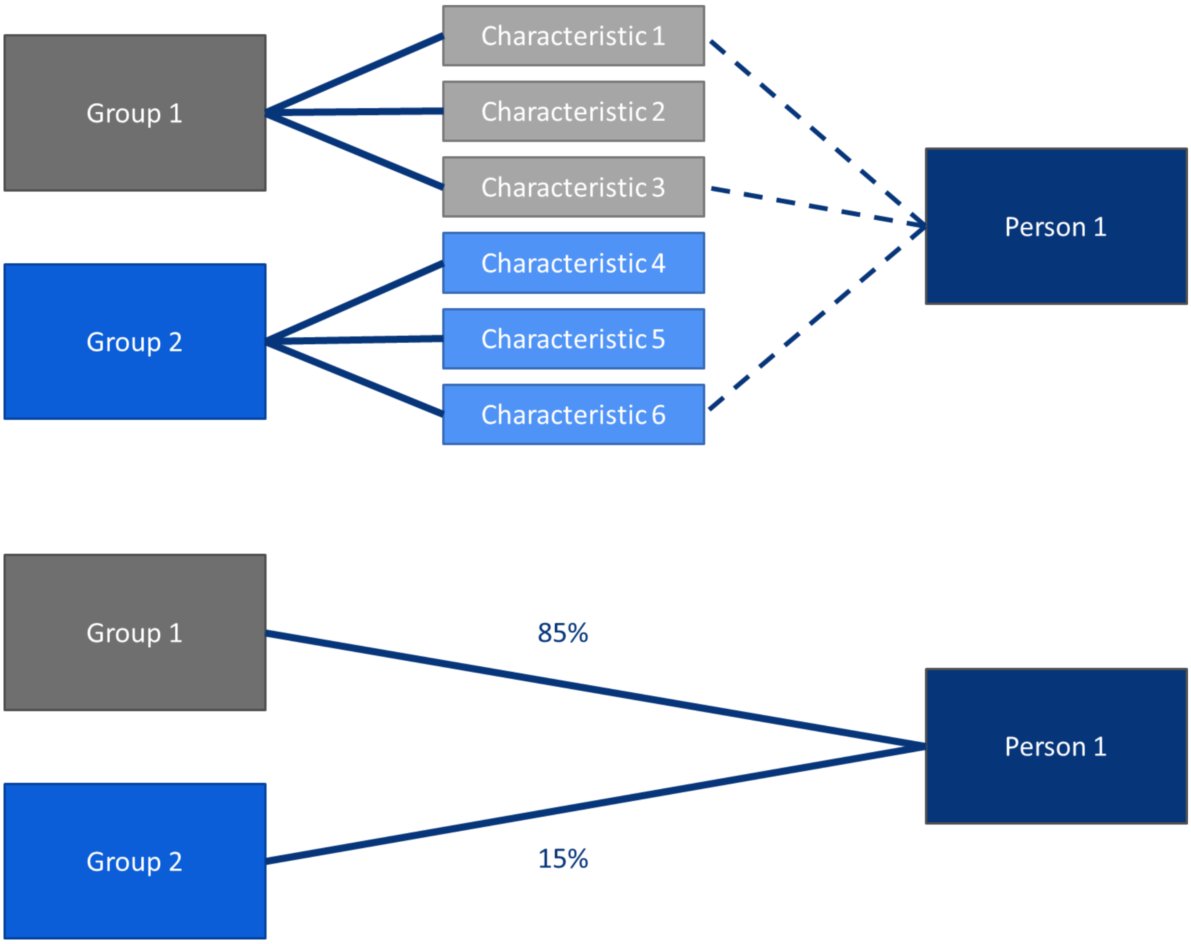
The classification model assumes that **each individual is a mixture of different groups** and that **each group is defined by a mixture of characteristics**. By finding the cluster of characteristics that define each group, the LDA can judge to what extent an individual is a member of different groups.

This unsupervised machine learning topic model clusters text into a specified number of topics, which has been used to analyse text in a variety of scientific settings, such as grouping medical studies (Wu et al., 2012). LDA has also been recently used in economics to analyse the transcripts of the Federal Reserve's policy-making decisions (Hansen et al., 2018), CEO behaviour (Bandiera et al., 2020), characteristics of political candidates (Lee, 2021), and political ideologies among voters (Draca & Schwarz, 2021).

**The LDA model assumes that each worker has characteristics that make them a mix of different groups. Each group is a combination of different characteristics that often appear together.**

**Error! Reference source not found.** provides a simplified example of an LDA model that shows one p erson's characteristics and potential membership of two groups.

Figure 22 Illustrative simplified example of the LDA model



On the left-hand side are two groups that are defined as combinations of characteristics that often appear together. In reality, it is not necessarily the case that each characteristic only appears in one group[23].

Further, the links between characteristics within each group may be relatively stronger or weaker, depending on how often they appear together. Characteristic 2 may not be particularly correlated with Characteristics 1 and 3 (i.e., not appear that often with Characteristic 1 and 3). As a result, if someone has Characteristic 2, it makes them more likely to be part of Group 1, but not as likely as if they had Characteristic 1 or 3.

There are six potential characteristics that a person could take in the model. On the right-hand side, Person 1 has Characteristic 1, Characteristic 3, and Characteristic 6.

One output of the LDA classification is the assignment of probabilities to individuals that they are a member of each group. For each individual the probabilities add up to one.

The LDA classification uses characteristics about each individual to update beliefs about how likely an individual is to be a member of each group. The prior belief is that there an individual is equally likely to be a member of each group, and these beliefs are updated with the combination of their characteristics. In Figure 22, Person 1 is estimated to have an 85% likelihood of being part of Group 1 and a 15% likelihood of being part of Group 2.

The **implementation of the LDA** can be summarised by the following process

3) **Preprocessing** of the data into 'worker biographies';
4) The user specifies the **number ($N$) of groups** that are to be identified. Further discussion about the optimal choice of groups is presented in Section7.4.3;
5) The **LDA identifies groups by finding clusters** across the $k$-dimensions of characteristics, where there are $k$ characteristics included in the data;
6) Each individual is **assigned a probability of being a member of each of the $N$ groups**;
7) **Interpretation** of the identified groups.

**Preprocessing of data**

The preprocessing of the data involves fewer steps than in other applications of the LDA[24], but still requires the conversion of labour market and characteristics data into a word text format. The LDA uses a 'worker biography' variable that includes information about the worker, so preprocessing would include converting continuous variables into binned/aggregated factors (such as '*age_25_29*' which is included in the biographies of workers between the age of 25 and 29). The robustness of this preprocessing is tested in Section 7.4.5.

[23] For example, the characteristic of being between the ages of 50 and 54 may be a characteristic that appears in multiple groups. In one group being between 50 and 54 may coincide with having children between the age of 15 and 19, while in another group being between 50 and 54 may coincide with certain occupations.

[24] When handling text data, the user specifies 'stop words' and punctuation that need to be removed, as well as lemmatizing words to ensure consistency across different forms of the same word. These specific steps are not required in this application of the LDA.

## 7.2 Results

### 7.2.1 Interpretation of results

Among others, there are three approaches used when characterising the groups that the LDA classification has identified.

1) What characteristics appear most often in each group?
2) What characteristics appear disproportionately more often in each group?
3) What proportion of all workers with that characteristic are in each group?

One characteristic may appear disproportionately often among workers in one group, but it may not be a particularly common characteristic and may not be a characteristic that many workers in that group have.

The LDA classification estimates probabilities that each individual worker is part of each of the ten groups. The word cloud visualisation focuses on the characteristics associated with each group, while the subsequent analysis of workers in each group assigns each worker to the group with which they are most likely to be a member[25].

### 7.2.2 Word cloud visualisation

Word clouds can be used to understand which characteristics are more common and/or are disproportionately more common in each group.

Font size (height) is used as a measure of which characteristics appear disproportionately more often in this group compared to across all groups.

The colour of the word is used to indicate how common the characteristic is across all groups. All characteristics (e.g., between the age of 18 and 24, living in Wales, working in sales and customer service occupations) are ranked by the proportion of those in all groups that have the characteristic. This ranking determines the colour of the word, with dark grey indicating that it is one of the least common characteristics, dark blue indicating that it is one of the most common characteristics, light grey to light blue indicating that it not particularly common or uncommon.

Figure 23 provides an example of a word cloud visualisation for Group 8 (the numbering of groups is arbitrary and solely used for labelling purposes), highlighting two characteristics: being between the age of 18 and 24, and living in Wales.

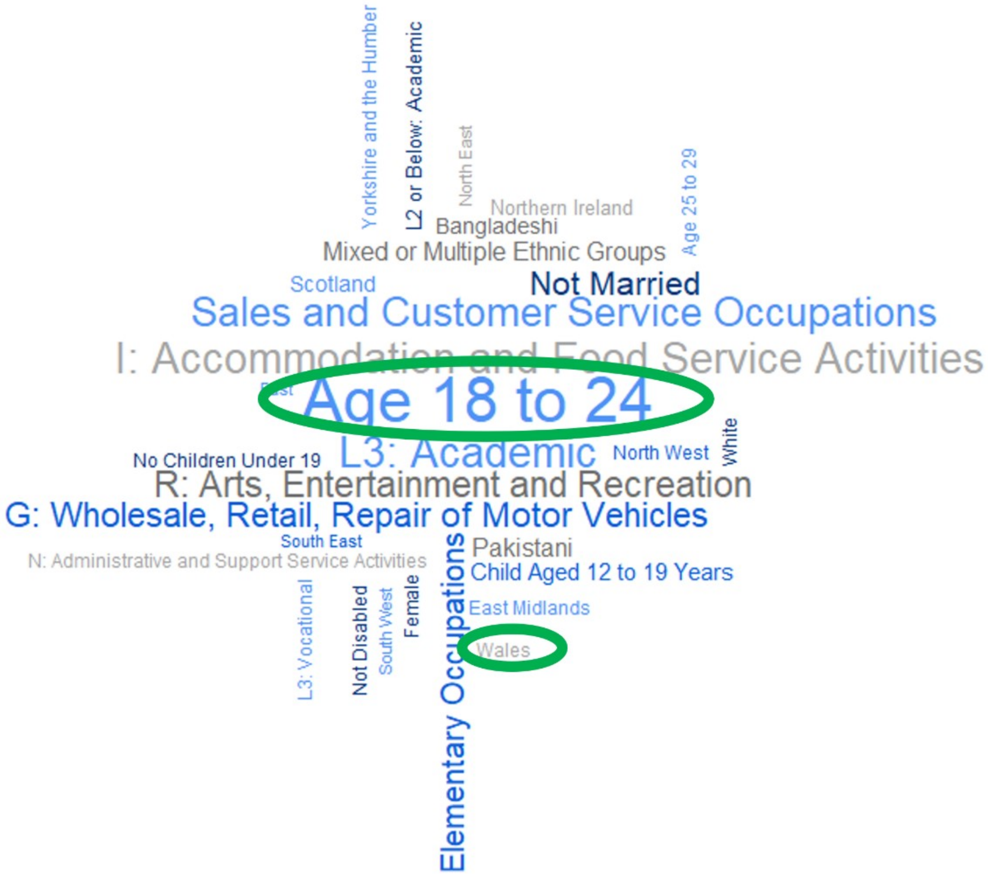
10% of workers across all groups are between the ages of 18 and 24, while 69% of Group 8 workers are between the ages of 18 and 24. As a result, workers in Group 8 are disproportionately likely to be between ages of 18 and 24, so the word cloud presents that characteristic in a larger font. In contrast, those in Group 8 are no more likely to live in Wales (5%) than workers across all groups (also 5%), so the characteristic is presented in a smaller font.

Figure 23 also provides an indication of how common the characteristics are across all groups using colours. 10% of workers across all groups are between the ages of 18 and 24 (i.e., a common

[25] As shown in Annex 7.4.4, the highest probability for each individual is on average 40%, four times the probability compared to random assignment of workers to the ten groups.

characteristic), whereas only 5% of workers across all groups live in Wales (a less common characteristic). As a result, the characteristic of being between the ages of 18 and 24 is presented in blue, whereas the characteristic of living in Wales is presented in grey.

**Figure 23 Example word cloud: Group 8**



***Source: London Economics analysis of LFS data***

### 7.2.3 Summary of the ten LDA groups

Table 4 presents a summary of workers assigned to each of the ten groups. These descriptions complement the descriptive interpretation of the associated word clouds, which are presented in the Annex.

The **group number listed in the first column is arbitrary** and solely used as a label for different groups. The second column reports key characteristics from each group. These include characteristics that are (disproportionately) common in that group. Due to the probabilistic nature of the LDA model[26] and the wide range of characteristics included, there are inevitably alternative possible interpretations of each group. Those that are presented in Table 4 are characteristics that distinguish a given group most from other groups and those that are common within the group.

[26] Having one characteristic does not deterministically rule out an individual being part of one group, unlike the decision tree analysis.

## Table 4 Groups identified by the LDA classification

| Group number | Description/Key characteristics | Minimum wage coverage |
|---|---|---|
| 1 | **Older workers** (60% are aged 50 or older, three times more likely to be disabled)<br>**Lower-level qualifications** (six times more likely to have no qualifications, 80% have Level 2 qualifications or below as their highest-level qualification or no qualifications)<br>**Elementary occupations** (43% work in elementary occupations, 33% work in sales and customer service occupations) | 31% |
| 2 | **Human health and social work activities and education** (58% work in human and health activities, 25% work in education)<br>**Ethnic minorities** (includes 22% of all those who are from 'other' Asian backgrounds and 24% of all those who are Black, African, Caribbean or Black British)<br>**Vocational qualifications** (four times more likely than the overall population to have a Level 3 vocational qualification as their highest-level qualification) | 28% |
| 3 | **Young London graduates** (six times more likely to live in London, five time more likely to be between 25 and 29 years old, 57% have a first degree or equivalent)<br>**Professional industries and occupations** (includes a quarter of all those working in financial and insurance activities and professional, scientific, and technical activities)<br>**Ethnic minorities** (includes a quarter of those who are Bangladeshi; Black, African, Caribbean or Black British; Chinese; Mixed or Multiple Ethnic Groups) | 4% |
| 4 | **Parents in their 30s with young children** (two thirds of the group are between 30 and 39 years old, five times more likely to have a child between the age of 0 and 4 years old – 58% of the group)<br>**Mostly male** (80% of the group are male)<br>**Vocational qualifications and skilled trades occupations** (twice as likely to have Level 3 vocational qualifications as their highest-level qualification, three times more likely to have Level 2 or below vocational qualifications as their highest-level qualification, four times more likely to be in a skilled trade occupation – around a quarter of the group) | 16% |
| 5 | **Manufacturing, construction, and transportation and storage** (five times more likely to work in manufacturing and construction industries)<br>**Older workers** (two thirds are aged 50 or older)<br>**Vocational qualifications** (includes 58% of all those who have Level 2 and 3 advanced, intermediate or trade apprenticeship qualifications) | 9% |
| 6 | **Parents between 30 and 44 years old** (90% are between 30 and 44 years old, 60% of the group have a child between the age of 0 and 4 years old)<br>**Asian ethnic minorities** (includes over a third of all Chinese and Indian workers, and around a quarter of Pakistani and other Asian background workers)<br>**Professional occupations and university degrees** (three times more likely to have a first degree or equivalent as their | 5% |

| Group number | Description/Key characteristics | Minimum wage coverage |
|---|---|---|
| | highest-level qualification, four times more likely to have a postgraduate degree as their highest-level qualification, three times more likely to be working in a professional occupation) | |
| 7 | **Public administration, real estate activities, and financial and insurance activities** (includes a quarter of all those working in financial and insurance activities, over a third of those working in real estate activities)<br>**Female workers in their 50s** (almost half of the group are in their 50s, 86% of the group are female)<br>**Lower-level qualifications** (39% of the group have Level 2 or below academic qualifications as their highest level) | 6% |
| 8 | **Young workers** (over two-thirds of the group are between 18 and 24 years old)<br>**Level 3 academic qualifications** (around a third of the group)<br>**Sales and elementary occupations** (two thirds work in sales and customer service occupations or elementary occupations) | 32% |
| 9 | **Parents in their 40s to early 50s** (almost 90% of the group are between the ages of 40 and 54, around three quarters of the group have children between the ages of 5 to 19 years old)<br>**Level 4 and 5 vocational qualifications** (four times more likely to have Level 4 or 5 vocational qualifications as their highest level)<br>**Associated professional and technical occupations and managers, directors, senior officials** (30% work in professional or technical occupations, 37% are managers, directors, or senior officials) | 5% |
| 10 | **Highly qualified** (44% of the group have a postgraduate qualification as their highest-level qualification, 40% of the group have a first degree or equivalent as their highest-level qualification)<br>**Professional occupations such as in education** (includes 40% of all those working in professional occupations, six times more likely to be working in education)<br>**Workers in their late 40s and 50s** (62% of the group are aged between 45 and 60 years old) | 2% |

Note: Minimum wage coverage indicates the proportion of those in the group who are earning below five pence above their relevant minimum wage. Group numbers are arbitrary and solely used as a label for groups. The rows highlighted in light blue are groups within which the minimum wage coverage is significantly higher than in other groups.

***Source: London Economics analysis of LFS data***

The characteristics that define each group (and distinguish workers from workers in other groups) range from occupation and education to age and ethnicity.

The results highlight the added value of categorising workers by more than one or two characteristics. For example, those from ethnic minority backgrounds appear disproportionately more often in Groups 2, 3, and 6, with very different minimum wage coverage between Group 2 and Groups 3 and 6.

### 7.2.4 Minimum wage groups

Although minimum wage status was not included as a characteristic in the LDA classification method, the identified groups highlight three groups where minimum wage coverage is significantly higher than other groups: **Group 1, Group 2, and Group 8** which are highlighted in light blue in Table 4.

**These groups are candidates for key types of minimum wage (or low pay) workers**, with the coverage of the minimum wage in these groups (28% to 32%) was around double of the next highest group (Group 4 with 16%) and coverage across all groups (14%).

The three potential minimum wage groups cover different types of workers across different characteristics. **Groups 1 and 8 differ in the age of the workers, while Group 2 contains disproportionately more workers from ethnic minority backgrounds compared to Groups 1 and 8.**

**The three groups cover workers with different levels of qualifications**. Group 8 contains younger workers with Level 3 qualifications who by virtue of their age may have soon after achieved higher-level qualifications, while Group 1 contains lower-level qualifications. Group 3 workers are more likely to have vocational qualifications (which is consistent with working in human health and social work activities and in education).

## 7.3 Evaluation

Unlike supervised machine-learning classification, where an outcome is being predicted (whether a worker is a minimum wage worker), the Latent Dirichlet Allocation cannot be evaluated in a similar manner. The results of the Latent Dirichlet Allocation are instead evaluated through sensitivity and robustness checks in the following section. These checks illustrate the extent to which the results are robust across changes in the model (for example, the number of groups specified).

## 7.4 Sensitivity and robustness checks

The number of groups is specified before the LDA classification is undertaken. As a result, it is important to evaluate the number of groups chosen. Although more groups may be more tightly defined by a set of shared characteristics, the set of many groups may be more difficult to interpret[27]. Further, the classification may suffer from overfitting: differences between groups may be driven by noise in the data.

On the other hand, specifying too few groups may result in multiple groups that have important differences being combined within a single group. That single group would not be particularly tightly defined[28].

### 7.4.1 Correlation between groups

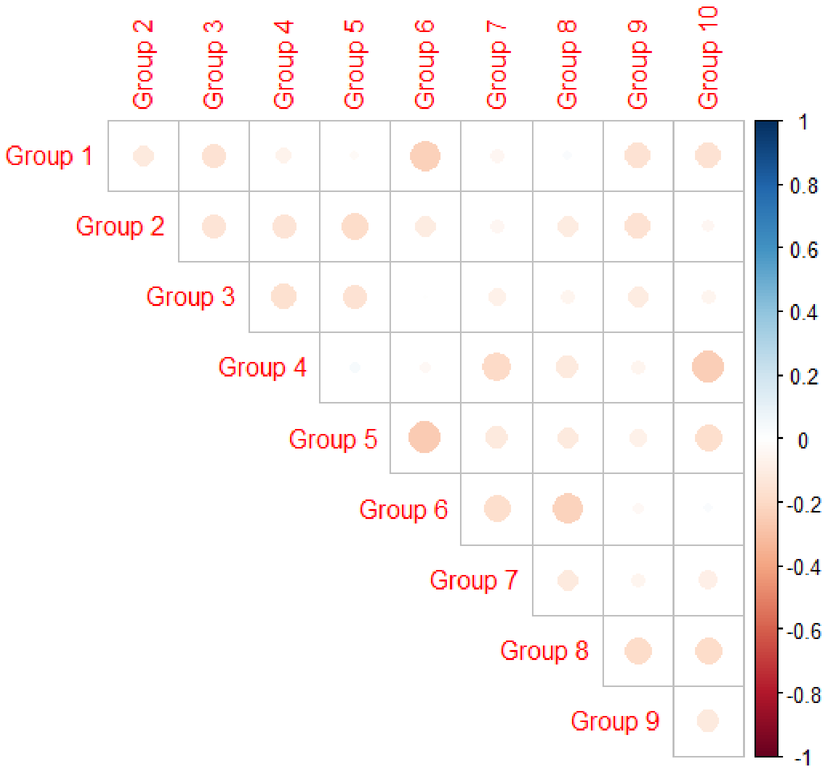
**One indicator that there are too many groups may be if there is a strong positive correlation between some groups**. If the probability of being one group is strongly positively correlated with being part of another group, this would potentially suggest that the division between the two groups may be arbitrary.

This is tested in Figure 24. Correlations between each group are presented in each cell, with the colour indicating the direction of the correlation (with darker colours indicating greater magnitude) while the size of the circle indicates the magnitude of the correlation (whether it is a stronger or weaker correlation, irrespective of direction).

[27] For example, specifying hundreds of groups may require further clustering across groups to be useful, as each individual group may only capture a small proportion of the labour market.

[28] One simplistic but illustrative example is a classification of two groups each defined by gender: male and female. While there may be significant differences in some characteristics between the two groups (such as working in different sectors and occupations), there are other characteristics (such as by education) that may also be important characteristics when segmenting the labour market.

Figure 24 Correlation between groups (workers' probabilities of being part of each group)



Note: The size of the circle indicates the absolute magnitude of the correlation between the probability of being part of each group, while the colour indicates the correlation.

***Source: London Economics analysis of LFS data***

Given that the probabilities for each worker sum to a total of one, it is expected that the correlations between probabilities of being part of each group will generally (but not necessarily all) be negative. If the probability of being in Group 1 increases from 40% to 60%, the probability of being in one of the other nine groups falls from 60% to 40%, although it is still possible that the probability of being in, for example, Group 2, increases from 5% to 10%.

It is notable that **there are no positive correlations between groups, which allays some concerns about groups overlapping and there being potentially too many groups specified**.

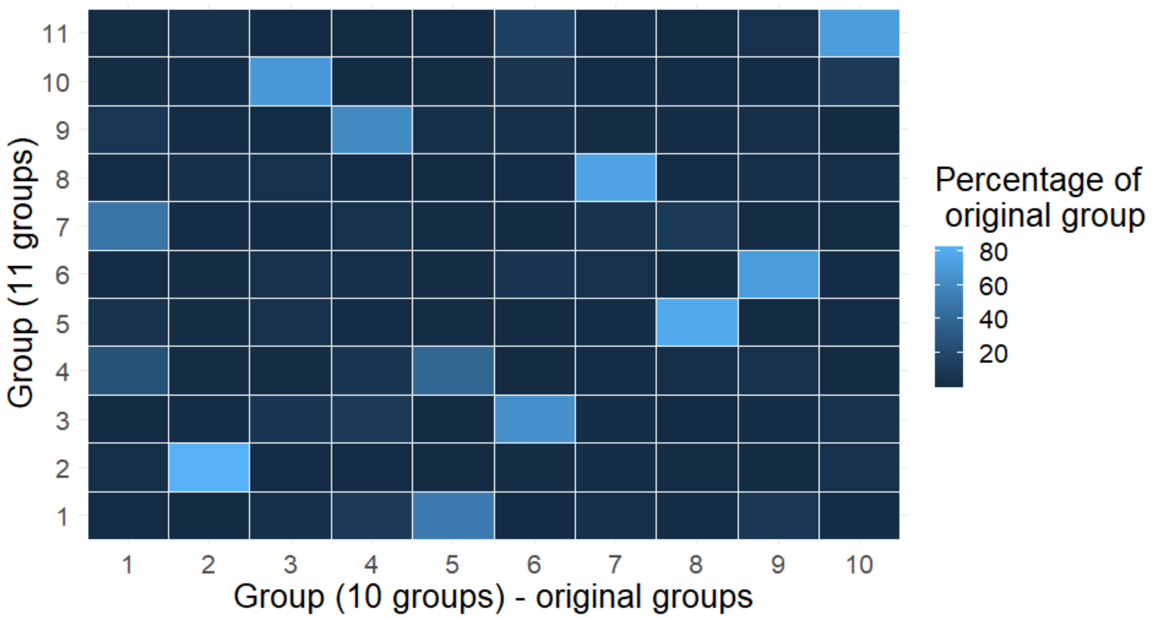
Some pairs of groups are more strongly and negatively correlated with each other. For example, Group 1 and Group 6 have a correlation coefficient of -0.23. As presented in Table 4, it appears that these groups of workers are quite different: Group 1 contains more older workers whereas Group 1 contains more workers in their thirties; Group 1 contains more workers with lower-level qualifications working in elementary occupations, while Group 6 contains more workers with university degrees working in professional occupations.

### 7.4.2 Stability of groups when changing the number of groups

To test the stability of the ten groups identified by the LDA classification, the LDA classification was rerun for eleven groups, and the heatmap above identifies the proportion of those in the original then groups that were assigned to each of the new eleven groups.

If the groupings found by the LDA classification were not stable, changing the number of specified groups would result in significant reshuffling of individuals between groups.

Figure 25 Stability of groups when changing the number of groups from ten to eleven



*Source: London Economics analysis of LFS data*

As illustrated in Figure 25, the ten groups identified by the LDA classification are relatively stable. The numbering of the groups is arbitrary, so stability can be interpreted as the extent to which the same individuals are grouped together both when ten groups are specified and when eleven groups are specified.

For example, the tile representing the proportion of those in group 2 (when ten groups were specified) that are in group 2 (when eleven groups were specified) is light blue, which indicates that a large proportion of those in group 2 remain in the same group as each other even when the number of groups changes.

Eight out of the original ten groups largely remain in the same group, with groups 1 and 5 from the original ten groups split in the formation of group 4 in the new eleven group specification.

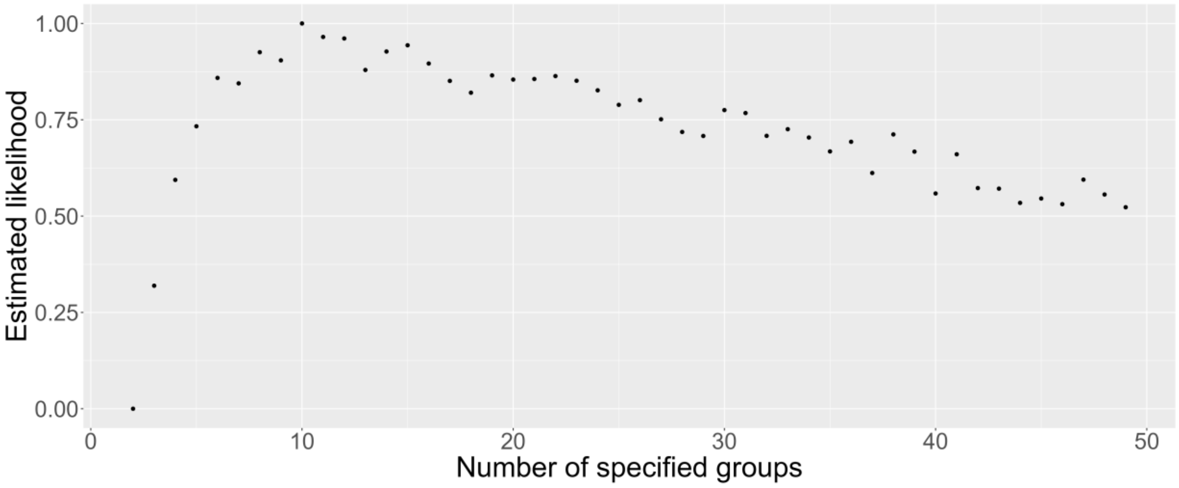
These results are consistent with the groupings not being particularly sensitive to changes in the number of groups. The addition of another group inevitably changes the composition of a few groups (groups 1 and 5 from the original ten groups), but the other eight groups are largely maintained.

### 7.4.3 Optimal number of specified groups

While the previous two tests provide an understanding of how robust the results are, an objective method can be used to identify the 'optimal' number of groups.

A method proposed by Griffiths & Steyvers (2004) is undertaken to identify the 'optimal' number of groups, who use a Bayesian method to estimate the likelihood of observing the distribution of characteristics across individual workers in the data for a given number of groups. This measure estimates the probability that the set of characteristics across workers is observed for a given number of groups, where $P(W|N)$ is the probability of observing data $W$ given $N$ groups. The LDA classification is rerun for each number of specified groups.

**Figure 26 Optimal choice of number of groups**



***Source: London Economics analysis of LFS data***

Figure 26 presents the estimated likelihood across different numbers of specified groups and normalises the likelihood scores between zero and one. **The likelihood initially increases as the number of specified groups increases but begins to level off at around ten groups**.

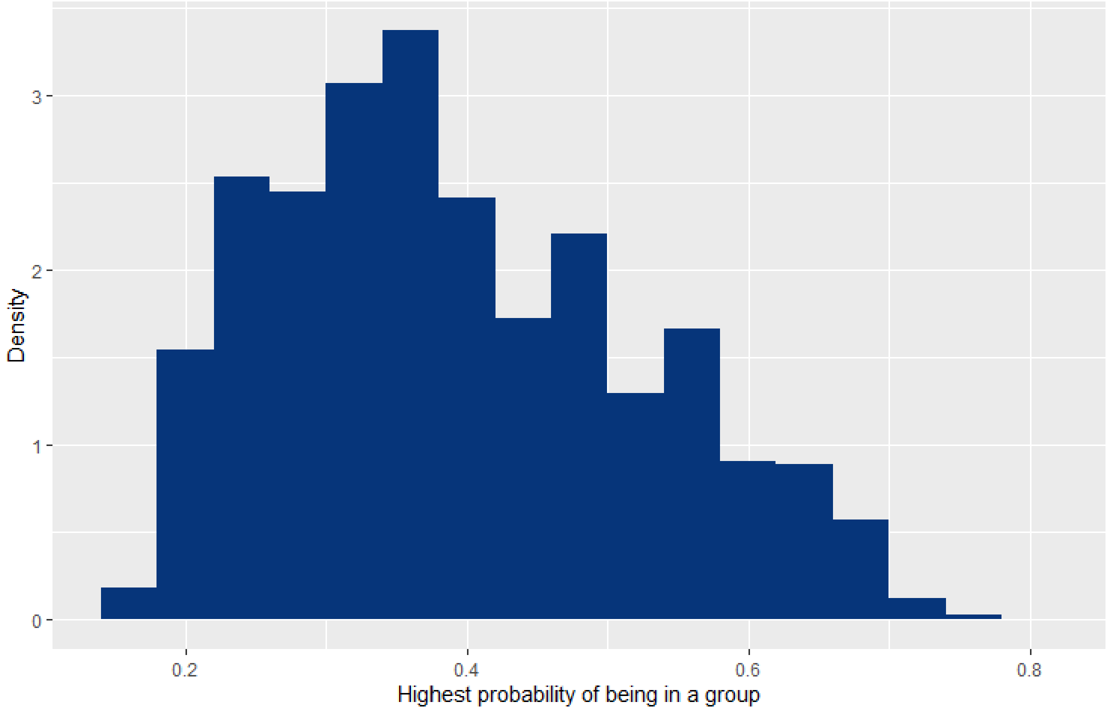
Beyond ten groups, the estimated likelihood decreases which is symptomatic of overfitting. If there are too many groups specified, this increases the chances that some of the groups identified by the LDA using the training data may not be found in other samples from the data (i.e., those **narrowly defined groups may be specific to the training data used and not generalisable**). This evaluation suggests that ten groups is an optimal number of groups.

As discussed by Griffiths & Steyvers (2004), **other practical considerations should be included when selecting the optimal number of groups** besides the number of groups suggested by the maximum likelihood method. These considerations may include **whether the identified groups are interpretable.** As mentioned in the discussion of the results, the ten groups identified are found to be intuitive groupings of workers, which further supports the specification of ten groups.

### 7.4.4 Highest probability of being in a group

Individual workers are assigned to the group that they are most likely to be a part of, which is on average 39.4%, with an interquartile range between 29.2% and 48.4%, which is also illustrated in Figure 27. The average highest probability of 39.4% is almost four times greater than the highest probability would be if there was random assignment, which illustrates the value of the information on characteristics in segmenting the labour force. If the highest probabilities were relatively low, this would suggest that the groupings were not particularly well-defined.

**Figure 27 Highest probability of being in a group across the ten groups identified by the LDA classification**

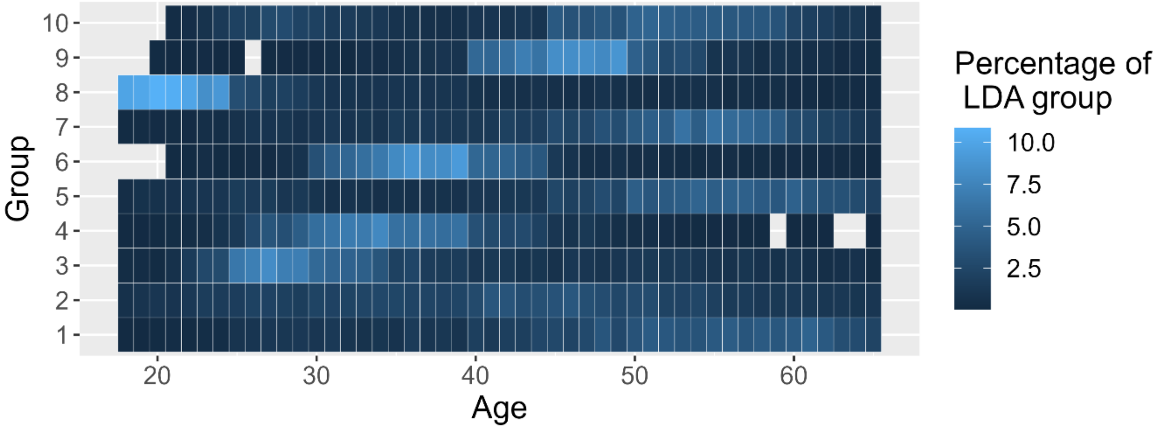


***Source: London Economics analysis of LFS data***

### 7.4.5 Definition of binned variables

One assumption made in the original analysis was how to categorise (into bins) characteristics such as age. In the case of ages, grouping by five-year bands (with the exception of those aged between 18 and 24 years old) may be too restrictive for the LDA classification if the distribution of ages within each group is very uneven. For example, if there are many 20-year-old workers in Group 8 but not many 21-year-old workers in Group 8, then grouping them together in one band would mask this heterogeneity.

**Figure 28 Distribution of ages within LDA groups**



Note: Each cell represents the percentage of each group that are contained with each single-year age group (for example, the percentage of workers in Group 8 who are 20 years old).

***Source: London Economics analysis of LFS data***

Figure 28 presents the distribution of workers' ages within each group. While some groups have pronounced distributions focusing on a particular age range (such as Group 8), others are less defined by a given age range (they may be defined by other factors such as occupation or education). In both cases, the distribution of ages within each group appears relatively smooth which suggests that banding ages together is not an overly restrictive step in the LDA classification. In addition, banding ages together, as mentioned in the methodology, improves the interpretability of the results.

# 8 Conclusion

This report uses two contrasting types of machine-learning classification techniques to identify groups of workers and groups of minimum wage workers.

The **decision tree** analysis highlights the importance of occupation and industry as predictors of whether an individual is a minimum wage worker or not. However, the decision tree analysis also identifies important differences across characteristics such as gender, age, educational qualifications within those occupation-industry groups. For example, the importance of being under the age of 21 with respect to minimum wage status varies is greater for some occupations than others. These differences may lead to a more targeted and specific understanding of transmission mechanisms through which changes to the minimum wage impacts labour market outcomes across groups. The decision tree classifications are able to predict minimum wage status using fifteen groups/dummy variables to a similar level of accuracy of models to other models that include many times the number of groups/dummy variables.

Further analysis tests the robustness of the decision tree findings by using methods that combine multiple decision trees: random forests and gradient boosted trees. These confirm the importance of occupation as a predictor of whether a worker is a minimum wage worker, followed by industry and qualifications.

The main analysis includes job characteristics (industry and occupation) which are key predictors. However, it is important to understand whether those not in work are likely to be minimum wage workers if they were employed (and so do not have industry or occupation information), as they may be impacted by changes in the minimum wage. Further analysis removes job characteristics and finds that qualifications and age are the most important predictors. However, the predictive performance significantly falls when job characteristics are not included.

The **Latent Dirichlet Allocation** classification identifies three potential minimum wage groups which can be generalised by the following characteristics:

4) older workers with lower-level qualifications working in elementary occupations,
5) workers with vocational qualifications working in human health and social work activities and education and are disproportionately likely to be from ethnic minority backgrounds, and
6) young workers in sales and elementary occupations who are yet to obtain higher-level qualifications.

Given the ability of the methods to identify intuitive groupings in an objective and systematic way as well as predicting minimum wage status, the groupings and techniques presented in this report could be used in further analysis that complements existing and ongoing minimum wage research (such as by the Low Pay Commission). For example, the groupings could be used to gain a more targeted understanding of how changes to the minimum wage (as well as other labour market events and interventions) may impact labour market outcomes differently for the different groups identified in this report.

Both the tree-based classification methods and the Latent Dirichlet Allocation provide new insights into the heterogeneity among workers. They are useful to policymakers in understanding what salient groupings exist among workers and minimum wage workers. For example, understanding different types of workers that are especially likely to be minimum wage workers provides potential explanations for how increases in the minimum wage may impact different types of workers more differently.

The tree-based classification methods also provide predictions for whether a worker, based on their characteristics, is likely to be a minimum wage worker. These predictions may be constrained by the data limitations (e.g., the number of characteristics available in the data) or in the case of this analysis by the dominance of occupation as a predictor of whether a worker is a minimum wage worker.

However, even when the decision tree's predictions may perform slightly worse than regression-based models, it is able to make those predictions using far fewer variables. For example, the decision tree, summarising the labour market in 15 groups (i.e., 15 dummy variables), performs only slightly worse than regression-based models that use over 100 variables.

Further analysis would explore other machine-learning methods, such as neural networks, that may provide better predictions for a given dataset. Other extensions would explore other data available that may have more characteristics available or have more reliable pay data (such as the ASHE and Census linked data). Specifically, the inclusion of past occupation could be a helpful predictor of whether a worker currently not employed would be a minimum wage worker if they were employed.

# References

Aitken, A., Dolton, P., & Riley, R. (2019). The Impact of the Introduction of the National Living Wage on Employment, Hours and Wages. *National Institute of Economic and Social Research (NIESR) Discussion Papers*, Article 501. https://ideas.repec.org//p/nsr/niesrd/501.html

Bryan, M., Salvatori, A., & Taylor, M. (2013). *The Impact of the National Minimum Wage on Employment Retention, Hours and Job Entry*.

Butcher, T., & Dickens, R. (2022). *Impact of the NLW using geographic wage variation* [Low Pay Commission in-house research report]. Low Pay Commission.

Cengiz, D., Dube, A., Lindner, A., & Zentler-Munro, D. (2022). Seeing beyond the Trees: Using Machine Learning to Estimate the Impact of Minimum Wages on Labor Market Outcomes. *Journal of Labor Economics*, *40*(S1), S203–S247. https://doi.org/10.1086/718497

Conlon, G., Lee, S.-M., Manly, L., & Patrignani, P. (2023). *Assessing the impacts of the reduction in the age of entitlement to the National Living Wage from age 25 to age 23*. Low Pay Commission. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1129902/Assessing_the_impact_of_the_reduction_of_NLW_age_of_eligibility_25_to_23.pdf

Cribb, J., Giupponi, G., Joyce, R., Lindner, A., Waters, T., Wernham, T., & Xu, X. (2021, December 9). *The distributional and employment impacts of nationwide Minimum Wage changes*. Institute for Fiscal Studies. https://ifs.org.uk/publications/distributional-and-employment-impacts-nationwide-minimum-wage-changes

Datta, N., McKnight, A., & Machin, S. (2021). Living wages and heterogeneous impacts by ethnicity, disability and gender. *Report for the Low Pay Commission. Centre for Economic Performance, London School of Economics and Political.*

Dickens, R., & Draca, M. (2005, June). *The employment effects of the October 2003 increase in the National Minimum Wage* (Working / Discussion Paper CEPDP0). (CEP Discussion Papers CEPDP0). Centre for Economic Performance (CEP), The London School of Economics and Political Science (LSE): London, UK. (2005); Centre for Economic Performance (CEP), The London School of Economics and Political Science (LSE). http://cep.lse.ac.uk/_new/publications/abstract.asp?index=2212

Dickens, R. & Lind. (2018). *The Impact of the Recent Increases in the Minimum Wage on the UK Labour Market: An Area-based Analysis*. Low Pay Commission.

Dickens, R., & Lind. (Forthcoming). *The Impact of the Recent Increases in the Minimum Wage on the UK Labour Market: An Area-based Analysis*.

Dickens, R., Riley, R., & Wilkinson, D. (2009). *THE EMPLOYMENT AND HOURS OF WORK EFFECTS OF THE CHANGING NATIONAL MINIMUM WAGE*.

Dickens, R., Riley, R., & Wilkinson, D. (2015). A Re-examination of the Impact of the UK National Minimum Wage on Employment. *Economica*, *82*(328), 841–864.

Druker, J., Stanworth, C., & White, G. (2002). *REPORT TO THE LOW PAY COMMISSION ON THE IMPACT OF THE NATIONAL MINIMUM WAGE ON THE HAIRDRESSING SECTOR*.

Dube, A. (2019). *Impacts of minimum wages: Review of the international evidence*.

Fidrmuc, J., & Tena, J. D. (2013). National Minimum Wage and Employment of Young Workers in the UK. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2286047

Georgiadis, A. (2006). *Is the Minimum Wage Efficient? Evidence of the Effects of the UK National Minimum Wage in the Residential Care Homes Sector*.

Georgiadis, A. (2021). *The Impact of the National Living Wage on the Adult Social Care Sector in England in the Context of COVID-19 Pandemic and Brexit*.

Giupponi, G., & Machin, S. J. (2018). Changing the Structure of Minimum Wages: Firm Adjustment and Wage Spillovers. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3170263

Hafner, M., Taylor, J., Pankowska, P., Stepanek, M., Nataraj, S., & Van Stolk, C. (2017). *The impact of the National Minimum Wage on employment: A meta-analysis*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR1807.html

Heyes, J., & Gray, A. (2001). Homeworkers and the National Minimum Wage: Evidence from the Textiles and Clothing Industry. *Work, Employment and Society*, *15*(4), 863–873. https://doi.org/10.1017/S0950017001008637

Lordan, G. (2019). People versus machines in the UK: Minimum wages, labor reallocation and automatable jobs. *PLOS ONE*, *14*(12), e0224789. https://doi.org/10.1371/journal.pone.0224789

Low Pay Commission. (2021). *National Minimum Wage Low Pay Commission Report 2021*.

Low Pay Commission. (2022). *Low Pay Commission Report 2022*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1129930/Low_Pay_Commission_Report_2022.pdf

Lucas, R., & Langlois, M. (2003). Anticipating and Adjusting to the Introduction of the National Minimum Wage in the Hospitality and Clothing Industries. *Policy Studies*, *24*(1), 33–50. https://doi.org/10.1080/01442870308040

Machin, S., Manning, A., & Rahman, L. (2002). *Where the Minimum Wage Bites Hard: The Introduction of the UK National Minimum Wage to a Low Wage Sector*.

Manning, A. (2016). The Elusive Employment Effect of the Minimum Wage. *Journal of Economic Perspectives*, *35*(1), 3–26. https://doi.org/10.1257/jep.35.1.3

Norris, G., Williams, S., & Adam-Smith, D. (2003). The implications of the national minimum wage for training practices and skill utilisation in the United Kingdom hospitality industry. *Journal of Vocational Education & Training*, *55*(3), 351–368. https://doi.org/10.1080/13636820300200234

Stewart, M. B. (2004). The employment effects of the national minimum wage*. *The Economic Journal*, *114*(494), C110–C116. https://doi.org/10.1111/j.0013-0133.2003.00200.x

Stops, M., Dolton, P., & Rosazza-Bondibene, C. (2012). *The Spatial Analysis of the Employment Effect of the Minimum Wage: Case of the UK 1999-2010*.

Vadean, F., & Allan, S. (2021). The Effects of Minimum Wage Policy on the Long-Term Care Sector in England. *British Journal of Industrial Relations*, *59*(2), 307–334. https://doi.org/10.1111/bjir.12572

Wilson, J., & Machin, S. (2004). Minimum wages in a low-wage labour market: Care homes in the UK. *The Economic Journal*, *114*(494), Article 494.

Winters. (2001). The Impact of the National Minimum Wage on the UK Thoroughbred Horseracing Industry. *Research Report to the Low Pay Commission*.

# Annex 1 Abbreviations used in the decision tree diagrams

## A1.1 UK SIC hierarchy

### Table 5 Section letters within the UK SIC hierarchy

| Section letter | Definition |
|---|---|
| A | Agriculture, Forestry and Fishing |
| B | Mining and Quarrying |
| C | Manufacturing |
| D | Electricity, Gas, Steam and Air Conditioning Supply |
| E | Water Supply; Sewerage, Waste Management and Remediation Activities |
| F | Construction |
| G | Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles |
| H | Transportation and Storage |
| I | Accommodation and Food Service Activities |
| J | Information and Communication |
| K | Financial and Insurance Activities |
| L | Real Estate Activities |
| M | Professional, Scientific and Technical Activities |
| N | Administrative and Support Service Activities |
| O | Public Administration and Defence; Compulsory Social Security |
| P | Education |
| Q | Human Health and Social Work Activities |
| R | Arts, Entertainment and Recreation |
| S | Other Service Activities |
| T | Activities of Households as Employers; Undifferentiated Goods- and Services-producing Activities of Households for Own Use |
| U | Activities of Extraterritorial Organisations and Bodies |

***Source: ONS Standard Industrial Classification (SIC) Hierarchy***[29]

[29] See https://onsdigital.github.io/dp-classification-tools/standard-industrial-classification/ONS_SIC_hierarchy_view.html for further information

## A1.2 Occupation classification

### Table 6 Sub-major groups of SOC2010

| Skill Level | Two-digit occupation code | Sub-major group within SOC2010 classification |
|---|---|---|
| Level 4 | 11 | Corporate managers and directors |
| | 21 | Science, research, engineering and technology professionals |
| | 22 | Health professionals |
| | 23 | Teaching and educational professionals |
| | 24 | Business, media and public service professionals |
| Level 3 | 12 | Other managers and proprietors |
| | 31 | Science, engineering and technology associate professionals |
| | 32 | Health and social care associate professionals |
| | 33 | Protective service occupations |
| | 34 | Culture, media and sports occupations |
| | 35 | Business and public service associate professionals |
| | 51 | Skilled agricultural and related trades |
| | 52 | Skilled metal, electrical and electronic trades |
| | 53 | Skilled construction and building trades |
| | 54 | Textiles, printing and other skilled trades |
| Level 2 | 41 | Administrative occupations |
| | 42 | Secretarial and related occupations |
| | 61 | Caring personal service occupations |
| | 62 | Leisure, travel and related personal service occupations |
| | 71 | Sales occupations |
| | 72 | Customer service occupations |
| | 81 | Process, plant and machine operatives |
| | 82 | Transport and mobile machine drivers and operatives |
| Level 1 | 91 | Elementary trades and related occupations |
| | 92 | Elementary administration and service occupations |

Note: The SOC2010 classification is the most recent occupation classification that allows for comparability across the entire sample used in the analysis, which covers the period from 2013Q2 to 2022Q3 inclusive (but exclusive of the furlough period as defined as 2020Q2 to 2021Q3 inclusive)

***Source: ONS, SOC2010 volume 1: structure and descriptions of unit groups***[30]

[30] See https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2010/soc2010volume1structureanddescriptionsofunitgroups for further information

## A1.3 Region, ethnicity, and qualification abbreviations

### Table 7 Abbreviations of regions

| Abbreviation | Description |
|---|---|
| E | East |
| EM | East Midlands |
| L | London |
| NE | North East |
| NW | North West |
| NI | Northern Ireland |
| SC | Scotland |
| SE | South East |
| SW | South West |
| WA | Wales |
| WM | West Midlands |
| Y | Yorkshire and the Humber |

Note: These are abbreviations used in the decision tree diagrams for UK regions

### Table 8 Abbreviations of ethnicities

| Abbreviation | Description |
|---|---|
| OA | Any other Asian background |
| Ba | Bangladeshi |
| Bl | Black, African, Caribbean or Black British |
| C | Chinese |
| I | Indian |
| M | Mixed or multiple ethnic groups |
| O | Other ethnic group |
| P | Pakistani |
| W | White |

Note: These are abbreviations used in the decision tree diagrams for ethnicities

***Source:* *ONS (link here)***

### Table 9 Abbreviations of highest level qualifications

| Abbreviation | Description |
|---|---|
| No | No qualifications |
| 2V | Level 2 vocational qualifications |
| 2Ac | Level 2 academic qualifications |
| 3V | Level 3 vocational qualifications |
| 3Ac | Level 3 academic qualifications |
| 45V | Level 4 and 5 vocational qualifications |
| FD | First degree or equivalent |
| OU | Other undergraduate |
| P | Postgraduate |

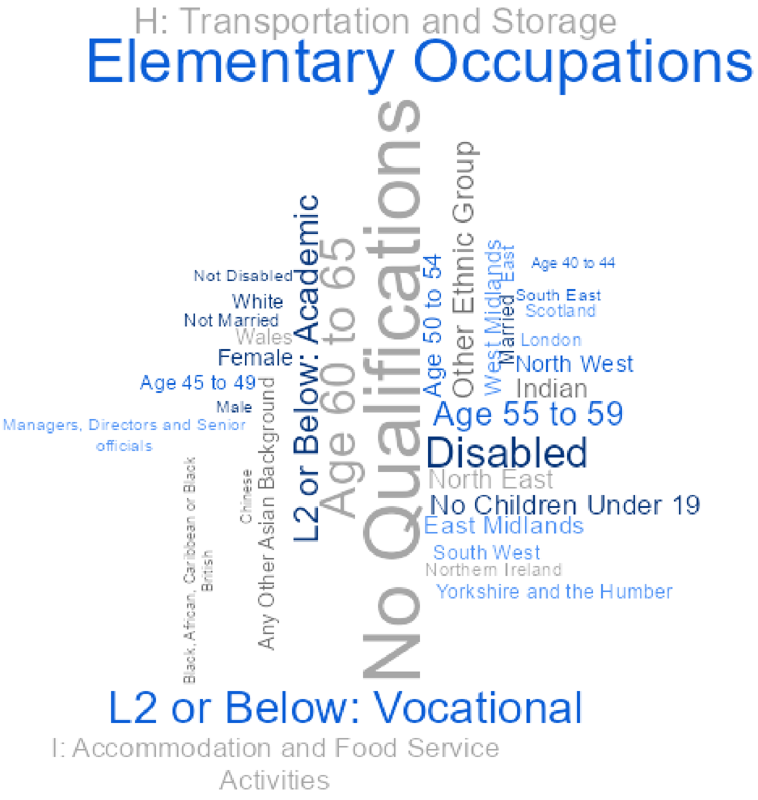
Note: These are abbreviations used in the decision tree diagrams for UK regions

***Source: ONS (link here)***

# Annex 2 Word clouds for LDA classification groups

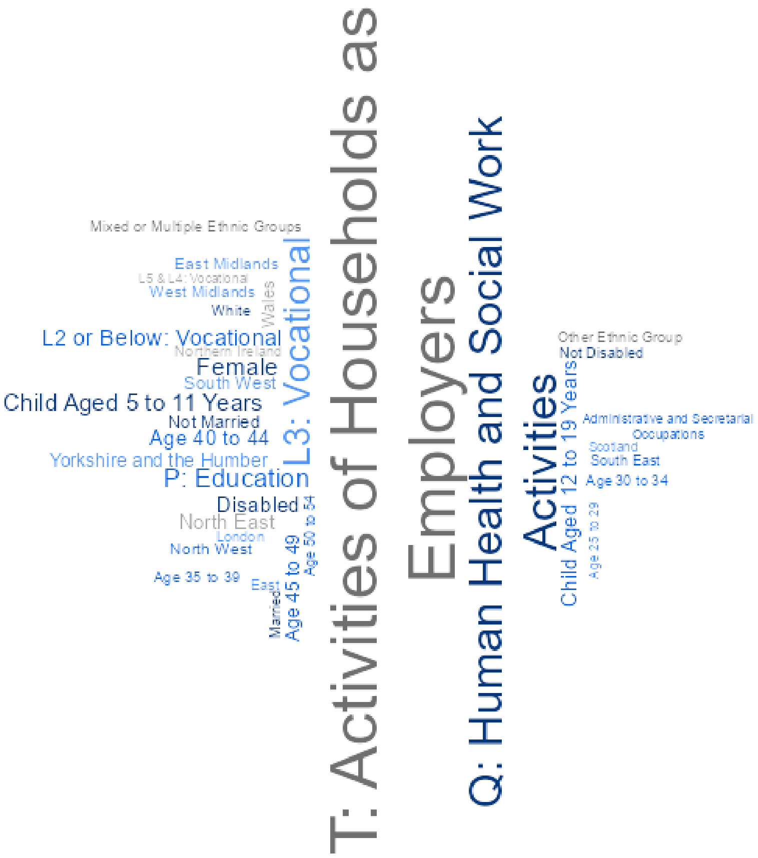
This annex presents the ten groups identified by the LDA classification method.

### Figure 29 Word cloud: Group 1



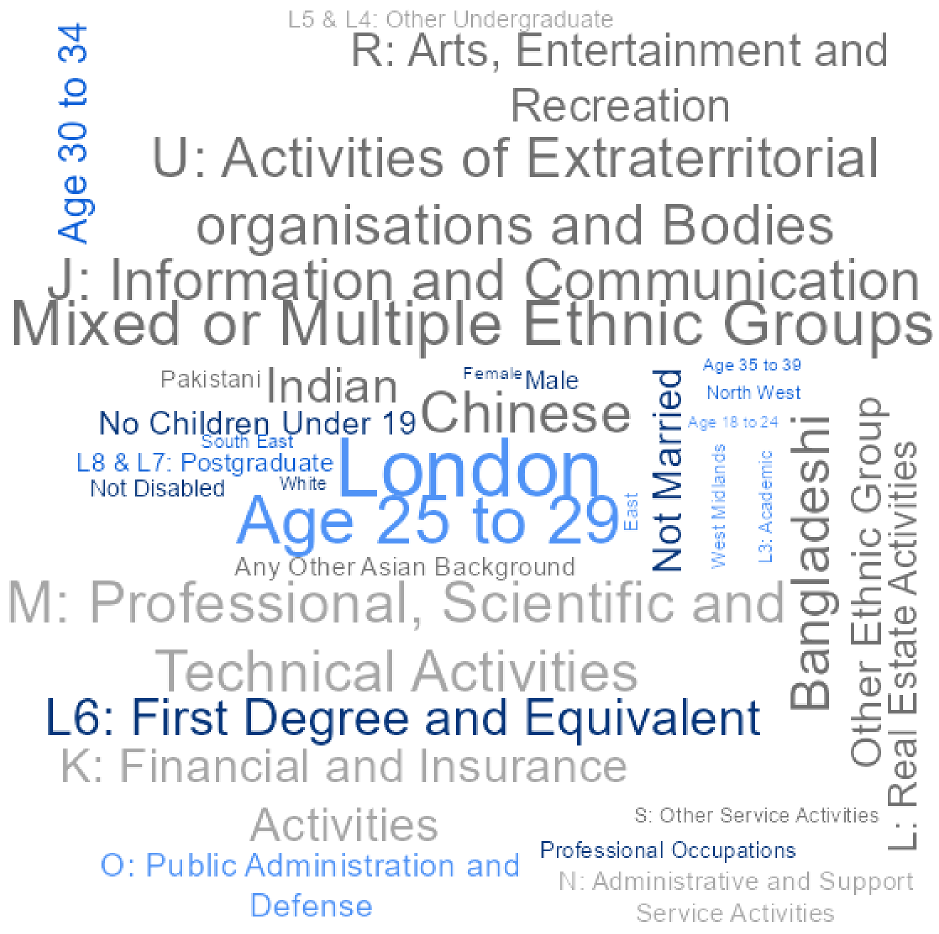
*Source: London Economics analysis of LFS data*

## Figure 30 Word cloud: Group 2



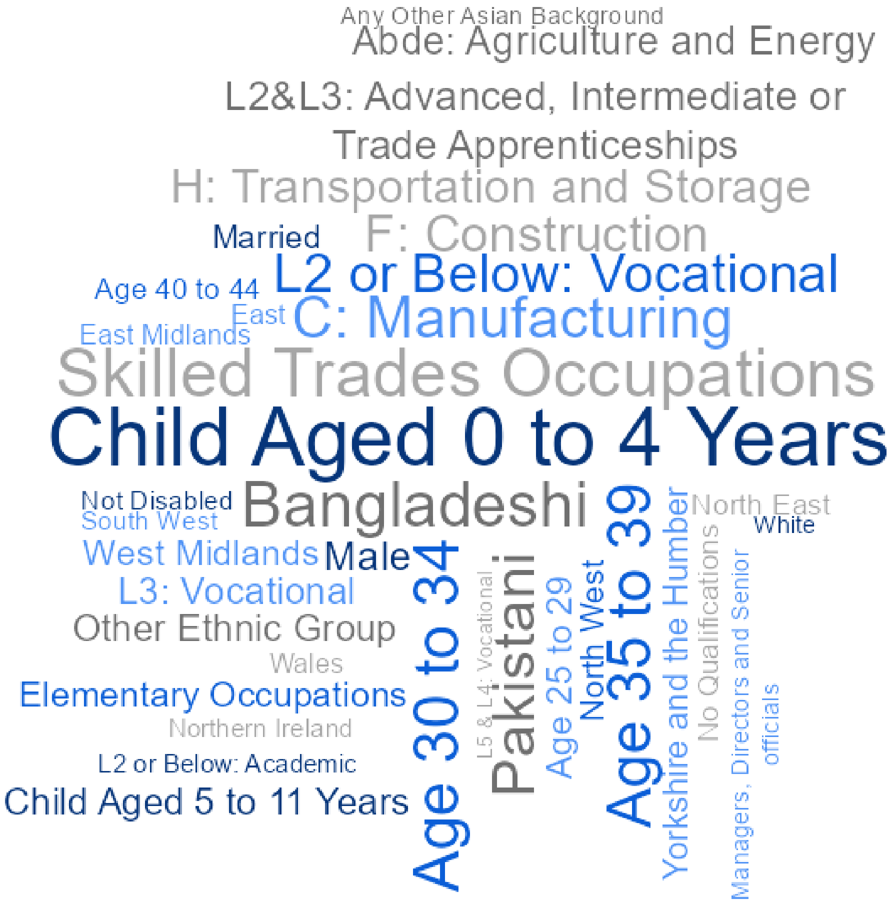
*Source: London Economics analysis of LFS data*

**Figure 31 Word cloud: Group 3**



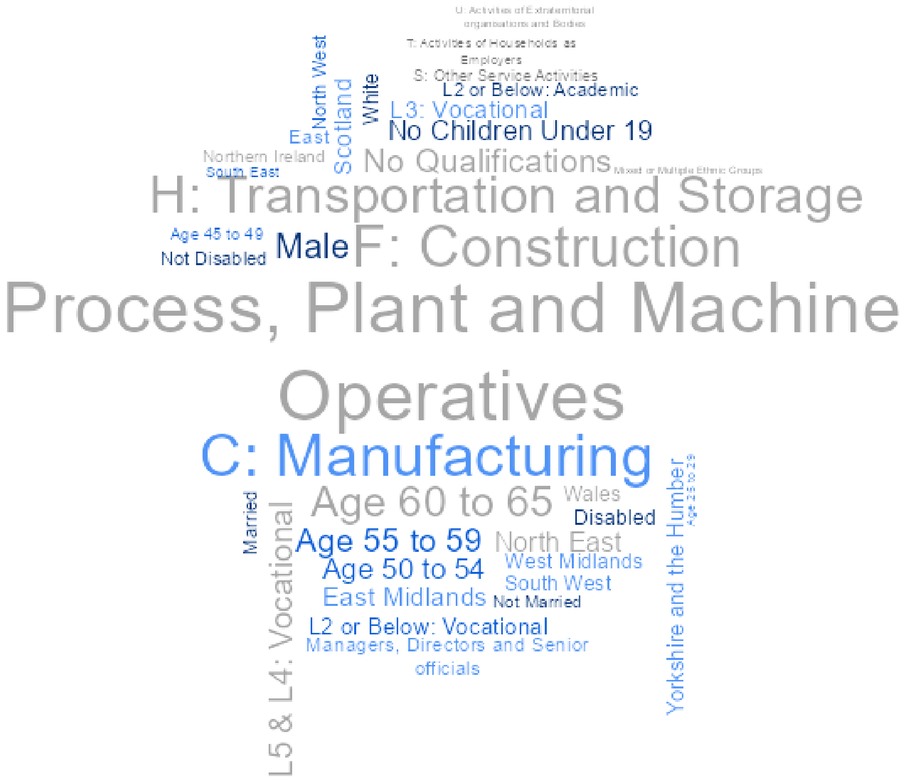
*Source: London Economics analysis of LFS data*

**Figure 32 Word cloud: Group 4**



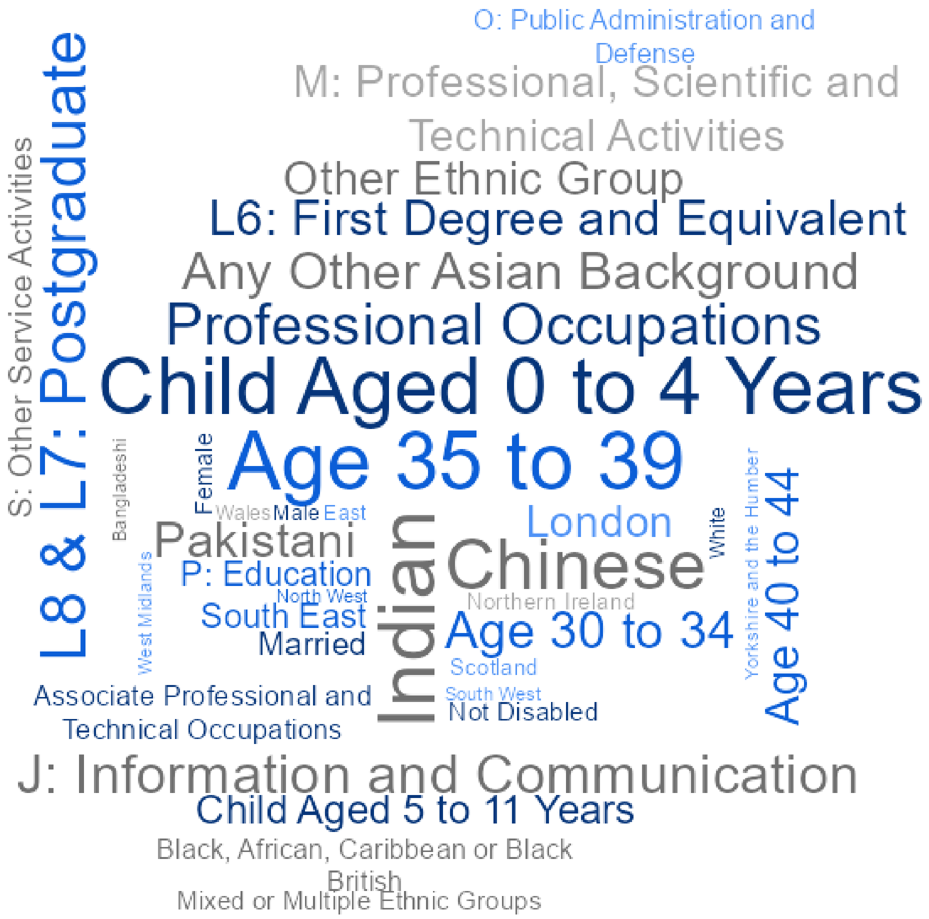
*Source: London Economics analysis of LFS data*

**Figure 33  Word cloud: Group 5**



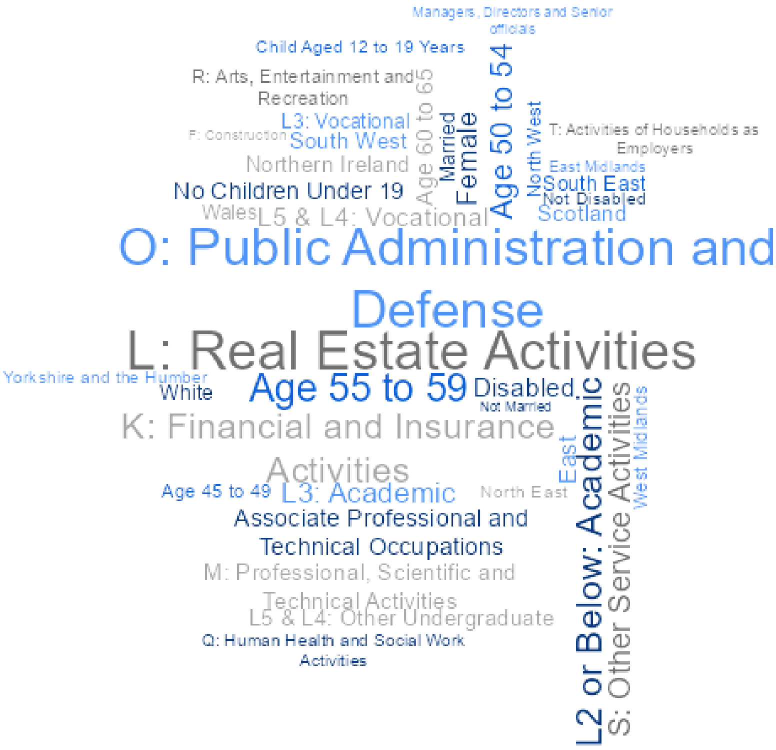
***Source: London Economics analysis of LFS data***

**Figure 34 Word cloud: Group 6**



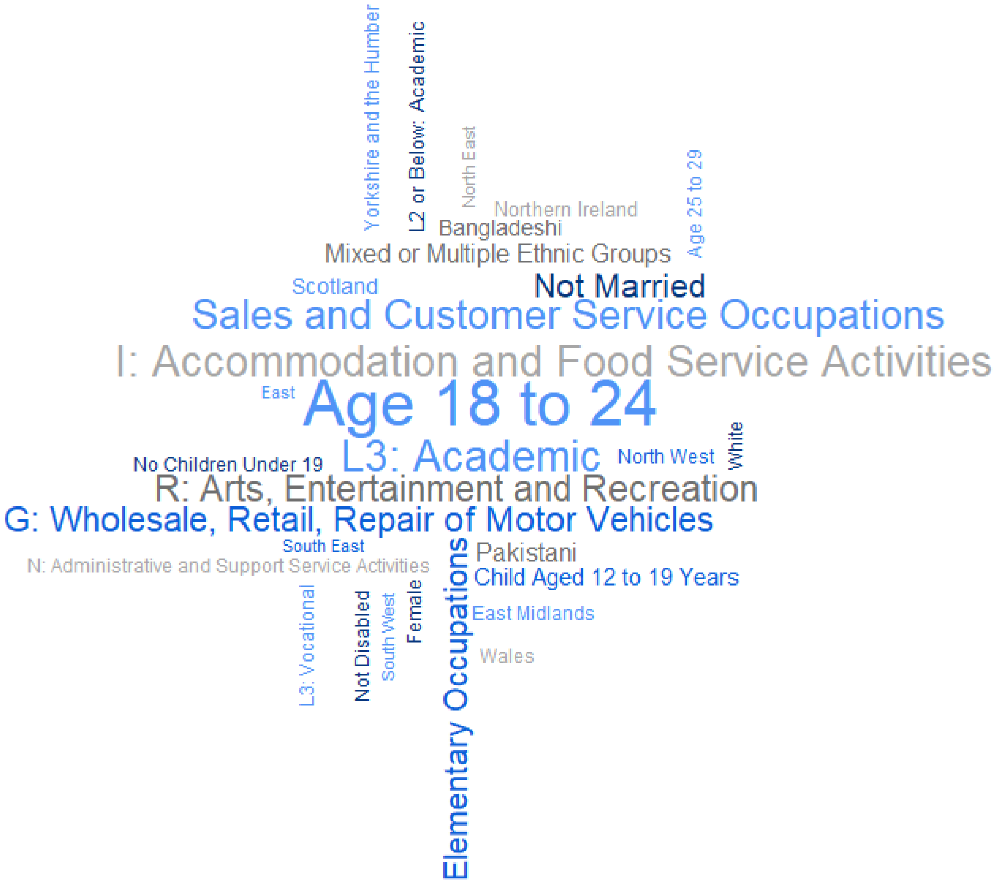
*Source: London Economics analysis of LFS data*

## Figure 35 Word cloud: Group 7



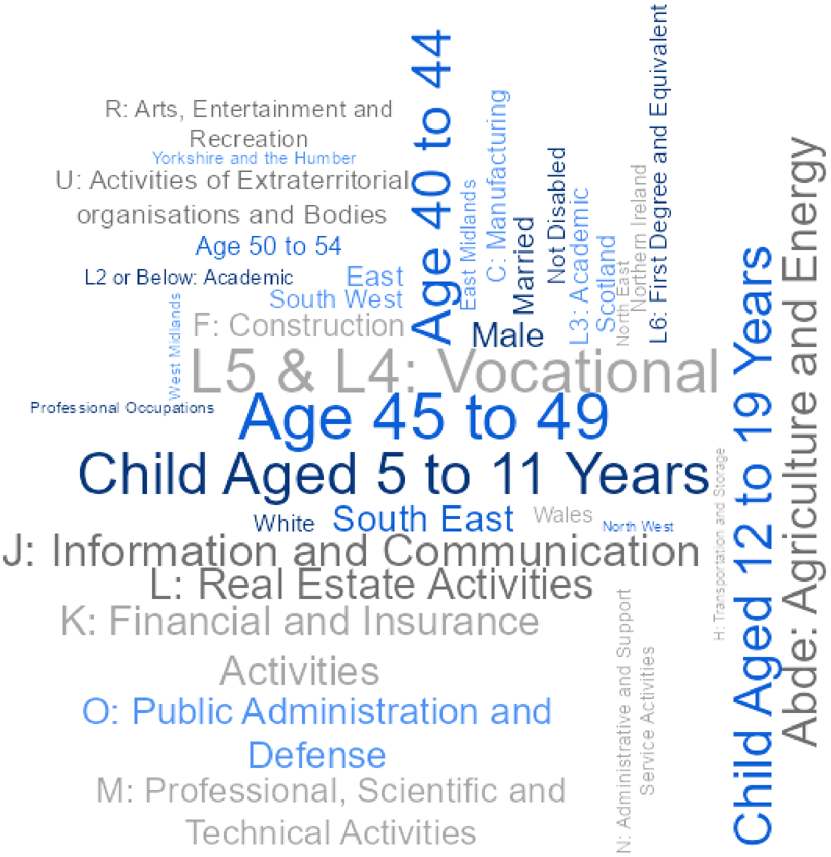
***Source: London Economics analysis of LFS data***

### Figure 36 Word cloud: Group 8



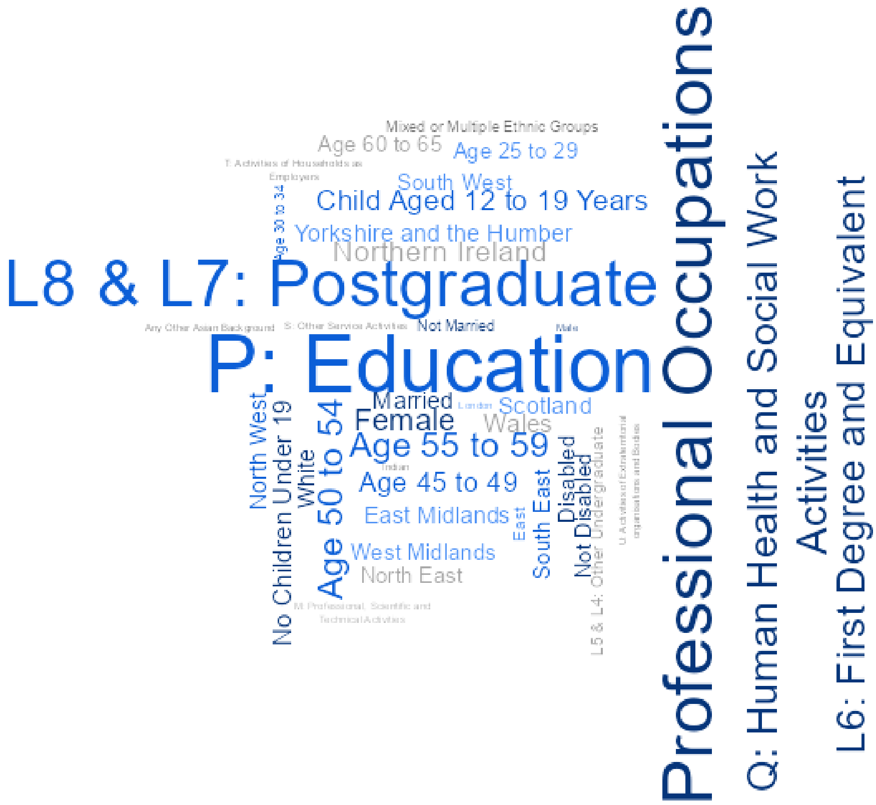
***Source: London Economics analysis of LFS data***

**Figure 37 Word cloud: Group 9**



***Source: London Economics analysis of LFS data***

### Figure 38 Word cloud: Group 10



*Source: London Economics analysis of LFS data*

LE
London
Economics

Somerset House, New Wing, Strand,
London, WC2R 1LA, United Kingdom
info@londoneconomics.co.uk
londoneconomics.co.uk
@LondonEconomics
+44 (0)20 3701 7700