Discussion Papers

DP/120/2022

Determinants of Labour Force Participation in Bulgaria: Empirical Evidence from Micro Data

Ventsislav Ivanov, Kristina Karagyozova-Markova,
Gergana Markova, Andrey Vassilev, Zornitsa Vladova

BULGARIAN NATIONAL BANK

BULGARIAN NATIONAL BANK

DISCUSSION PAPERS
DP/120/2022

Determinants of Labour Force Participation in Bulgaria: Empirical Evidence from Micro Data

Ventsislav Ivanov, Kristina Karagyozova-Markova, Gergana Markova, Andrey Vassilev, Zornitsa Vladova

2022

ISBN 978-619-7409-26-0 (online)

Send your comments and opinions to:
Publications Council
Bulgarian National Bank
e-mail: BNB_Publications@bnbank.org
Website: www.bnb.bg

Contents

Abstract: This paper explores the factors influencing labour market participation decisions in Bulgaria based on an anonymised micro dataset from the Labour Force Survey over the period from 2000 to 2019, provided by Eurostat. An important benefit of the data is that it contains a number of specific individual and household characteristics with potential relevance for labour supply decisions. For the analysis of this rich dataset we employ complex modelling techniques in order to address important statistical issues such as a large number of potential predictors as well as possible non-linear relationships. The techniques include adaptive lasso and adaptive group lasso from the statistical domain and various methods from the machine learning literature. The results confirm the stylised fact that the probability of participation is hump-shaped with respect to a person's age. Our main finding, which is robustly supported by most of the models, is that educational attainment plays a key role for labour force participation. A common finding is also the relevance of specific household characteristics such as the number of household members that are employed or inactive, although the causal association between these household characteristics and labour supply needs further investigation. We also obtain empirical evidence that the expansionary phase of the business cycle is positively associated with the probability of being part of the labour force.

Keywords: *labour force survey, micro data, statistical variable selection, machine learning*

JEL classification: C25, C52, J20, J22

Ventsislav Ivanov, Economic Research and Forecasting Directorate, Bulgarian National Bank, Ivanov.V@bnbank.org

Kristina Karagyozova-Markova, Economist, Directorate General International and European Relations, European Central Bank, Kristina.Karagyozova-Markova@ecb.europa.eu

Gergana Markova, worked in the Economic Research and Forecasting Directorate of the Bulgarian National Bank in the period 2016–2020, geri.mark@gmail.com

Andrey Vassilev, Dill Advisory, andrey.vassilev@dilladvisory.com

Zornitsa Vladova, Economic Research and Forecasting Directorate, Bulgarian National Bank, Vladova.Z@bnbank.org

1. Introduction

Bulgaria faces serious demographic challenges since the beginning of the 1990s. Negative demographic trends are associated with a contraction of the labour force and result both from the long-term decrease in the population and the gradual process of its ageing. According to the 2021 Ageing Report of the European Commission, over the period 2019–2070 the total population in Bulgaria is projected to contract by 27.8 per cent, which represents one of the sharpest declines among EU member states. As a whole for the EU, over the same period baseline demographic projections point to a decline of the population by 5.2 per cent. Furthermore, by 2070 working age population in Bulgaria (20–64 years) is forecast to shrink by 38.5 per cent compared to its level in 2019. In the medium term, the shrinking labour force may be expected to have a relatively moderate restraining effect on the country's potential economic growth. However, in the long run, the negative economic effects of these developments in the labour market are expected to become increasingly more significant. Changes in the number and age structure of the population are expected to affect household preferences for consumption, savings and labour supply, as well as all other factors of production, thus affecting long-term economic growth and price processes. Negative demographic trends also create potential risks for the long-term sustainability of public finances.

The size of the labour force depends both on the number of the total population, which according to Eurostat data has been following a pronounced downward trend with an overall decline by 20.7 per cent over the period 1990–2020, and on the activity of working age population. Compared to the average EU level, Bulgaria has a relatively high share of people outside the workforce and a high share of persons between 15 and 29 years of age who do not work or study. Studying the various determinants of labour force participation of the different groups of the population presents an interesting area of research which can shed light on the potential role of policies that could help increase labour supply in the country.

There are many factors with potential influence on the decision of individuals to participate in the labour market. These factors include personal and household characteristics, the stage of the economic cycle, structural changes in the economy, technological improvements as well as the design of the social security system and the overall institutional setting. Gender and age, as well as the decisions of individuals on the scope and duration of education, family formation and number of children have been identified in existing literature as important individual characteristics shaping the labour supply decision. A well-established finding in a number of studies is that wage growth and

the expansionary phase of the economic cycle have a positive impact on the participation in the labour market, while an overly generous social security system can lead to reduction of labour supply. An example of structural factors affecting labour supply is the long-term trend of expansion of the services sector which has likely contributed to the increasing share of females in the labour force. A common finding in the labour supply literature is also that access to childcare and greater flexibility in work arrangements tend to enhance female labour supply.

The purpose of this study is to establish the importance of the different socioeconomic factors (mainly personal and household characteristics) influencing the decision of individuals to participate in the labour market. We use annual anonymised micro data from the Labour Force Survey (LFS) for Bulgaria provided by Eurostat for the period 2000–2019. The dataset contains various individual and household characteristics that might be considered relevant for labour force participation. For the purpose of the analysis, we use descriptive and modelling techniques to look into the potential determinants of labour market participation. The modelling tools employed can broadly be grouped into two groups – models based on statistical methods that have certain desirable asymptotic properties and models based on methods from the machine learning literature. Among the statistical methods we apply the adaptive lasso and the adaptive group lasso to logistic regression as some of the commonly used linear shrinkage methods that enable us to establish only those predictors that have nonzero coefficients. Our choice of machine learning techniques is driven by the motivation to test several models enjoying a reputation for high predictive performance and combine them with procedures which aid the interpretation of model results. The decision to employ various selection and modelling techniques in the analysis is motivated by several factors. First, we aim to include a large initial number of potential explanatory variables, which can be achieved with the substantial volume of our micro dataset, but requires the use of empirical approaches that allow consistent variable selection. Second, the application of a number of models ensures greater reliability of the results and increases the robustness of the conclusions about the relevance of the driving factors for labour force participation in Bulgaria. Finally, the use of machine learning methods brings in additional gains as regards the ability to account for the impact of possible non-linearities in the analysed relationships. To the best of our knowledge, this is the first study to analyse micro data from the LFS for Bulgaria, using such a rich set of modelling techniques. In this regard, the article can serve as a reference for future research on labour market trends in Bulgaria based on the same dataset.

The results from the analysis confirm the stylised fact that the probability of participation is hump-shaped with respect to a person's age. A main finding, which is robustly supported by most of the modelling approaches, is that educational attainment is a major determinant of labour force participation. Higher level of education significantly increases the probability of being economically active. Furthermore, some evidence is also found for the specific field of education as an additional contributing factor. A common finding from the application of the different modelling techniques is also the relevance of specific household characteristics such as the number of household members that are employed or inactive, although the causal association between these household characteristics and labour supply needs further investigation. We also obtain empirical evidence that the expansionary phase of the business cycle is positively linked with the probability of being part of the labour force.

The paper is organised as follows. In Section 2 we review the empirical literature on the determinants of labour force participation. In Section 3 we present a description of the micro data in the LFS for Bulgaria with a focus on stylised facts on labour force participation in the country. Section 4 presents the chosen approaches to modelling the determinants of labour force participation and discusses the results obtained. Section 5 concludes the paper.

2. Literature Review

The neoclassical labour economics can be considered as the main theoretical foundation of the modern empirical literature on labour supply. Neoclassical labour economics applies the basic consumer demand theory as shown by papers such as Mincer's 1962 paper on labour force participation of married women (Mincer, 1962) and Becker's works respectively on the theory of the allocation of time (Becker, 1965) and the theory of social interactions (Becker, 1974).

Wages are assumed to be a key determinant for an individual's participation in the labour market in many theoretical models on labour supply. As Mincer (1962) notes, the response of working hours supplied to variations in the wage rate entails a positive substitution effect and a negative income effect and the overall effect on working hours cannot be determined a priori. Unlike standard regression models that estimate the wage elasticity of working hours or the so-called "intensive margin" of labour supply, there has been no established specification for modelling labour force participation decision or the "extensive margin" of labour supply in the theoretical and empirical literature. According to an overview of the relevant literature (Strauch et al., 2008), both types of labour supply decisions are driven by individual preferences and budget

constraints over the life cycle of a person, while institutions and structural conditions also play a significant role. The participation decision is mainly influenced by factors such as the tax system and the generosity and duration of unemployment benefits as well as by regulations affecting the flexibility of working hours (*e.g.* childcare and parental leave). The decision regarding the number of additional hours of work is primarily driven by the motivation for higher current or expected net income. Supplementary factors that influence labour supply behaviour are the family composition and stage in family cycle (*e.g.* marital status, number and age of children, joint household decisions taken within the family) and decisions relating to the life-cycle of a person (how many years to participate in the labour market, investment in human capital, *etc.*). Other important factors that can be included as some of the main determinants of the duration of a person's participation in the labour market are the characteristics of the national social security systems and the pension systems, which affect for example the incentives for early retirement and the statutory retirement age. The role of business cycle conditions as a potential determinant of labour force participation has also been studied extensively. Some authors also emphasise that labour force participation decisions are usually being characterised by persistence. For example, Clark and Summers (1982) point out that past work experience and accumulation of human capital along with high separation costs tend to raise the probability of subsequent employment once being employed, while at the same time developing household-specific commitments may reduce the attractiveness of engaging in work activities.

Pencavel (1986) cites Bowen and Finegan's significant contribution to the empirical literature from 1969 that explores factors affecting labour force participation across men with linear probability estimates based on micro data from the US census of population. The two authors find that years of schooling has a positive influence on the probability of being in the labour force. Killingsworth and Heckman (1986) also refer to Bowen and Finegan's study, emphasising that the educational attainment of a woman is strongly associated with higher participation probability. Being married or having a large amount of income other than own earnings were related to lower participation probability. A clear inverted U-shaped-relation between the probability of participation and age was additionally reported along with a negative effect of the presence of children (particularly under the age of six) on the probability for a married woman to participate in the labour market.

Recent studies confirm most of these early findings about the role of personal characteristics. Vlasblom and Schippers (2004), Bachmann et al. (2010) and Grigoli, Koczan and Topalova (2018), all estimating logit models based on LFS micro data, find statistically significant positive effects of higher education

on labour force participation. The results from the first two studies also show negative effects of having children on female labour force participation. These studies additionally reveal that the presence of younger children has a stronger negative effect on female participation decisions. At the same time, the negative impact of children is gradually mitigated as they get older.

A number of studies explore the influence of household characteristics on labour force participation in a much more significant detail and furthermore highlight potential econometric issues in most of the applied models in the economic literature, including the standard probit/logit models. Angrist and Evans (1998) analyse the effect of childbearing on the labour supply of their parents, stressing the endogenous nature of the fertility variable because of the potential joint determination of both fertility and labour supply. The authors employ data on parental preferences for a mixed sibling-sex composition to construct instrumental variables estimates, reaching the conclusion that children do lead to a reduction of female labour supply, although the ordinary least squares estimates appear to overestimate the causal effect of children. In a similar vein, Del Boca, Pasqua and Pronzato (2008) investigate the joint decision by women on work and childbearing with a bivariate probit model, while controlling for factors such as personal characteristics, childcare system, parental leave arrangements, family allowances and part-time work opportunities. Their results suggest that the different social policies across European countries explain a non-negligible percentage of the differences in women's labour market participation across these countries. Women with secondary and tertiary levels of education are significantly more likely to be in work and to have a child than women with primary education, whereas the presence of children of any age has a negative effect on the probability of working and decreases with the age of the child.

Some of the other personal and household characteristics that have been explored in the literature as relevant for the labour supply decision include gender, birth cohort propensities and elderly care responsibilities. A number of studies, both descriptive and model-based, have long documented that men have higher labour force participation rates than women. A joint study by the International Labour Organisation and UN Women (Azcona et al., 2020) finds that women's participation in the labour market varies more than men's across household types and that being part of a couple, especially with young children is associated with lower participation rates for women and higher rates for men. Cohort or birth-specific effects, particularly relevant for female labour force participation, have also been discussed extensively in the literature. Fallick and Pingle (2006) emphasise the long-term increase in female labour force participation as evident by entering cohorts of women with higher participation

rates for their ages than cohorts preceding them. The authors associate these developments with factors such as evolving tastes, reproductive technology, wealth, education, social attitudes, and the development of the retirement, welfare, and financial systems. Elderly care responsibilities within the family have not received as much attention as other household characteristics. Using a standard probit model, Cipollone, Patacchini and Vallanti (2013) find a negative and significant impact of elderly care responsibilities on female labour force participation based on individual household data for 15 EU countries.

The literature exploring the effects of different policy reforms (*e.g.* tax reforms and pension system reforms) and changes in other institutional characteristics of labour markets on labour supply is vast. Blundell (1995), who analyses tax reforms from the 1980s in the United Kingdom, emphasises the importance of micro-simulations that allow correct modelling of the tax and benefit schedule faced by an individual, and additionally enable responses to wage and income changes to vary between individuals with different demographic characteristics and economic conditions. One of his findings is that married women are the group with strongest responses to tax reforms, while the author also stresses the importance of studying the potential behavioural effects of tax reforms in a wider household context.

Another broad strain of literature deals in a more general aspect with the role that social policies and institutional factors play for participation in the labour market, especially in the case of female participation. A common finding in most empirical studies in this vein of literature is that available access to childcare, maternity leave policies and greater flexibility in work arrangements tend to enhance female labour supply. Based on a logit model for the participation rate of prime-age women, Genre, Gomez-Salvador and Lamo (2005) find that maternity leave stimulates labour supply with a statistically significant impact, however this effect turns negative after around 10 months. In line with other authors' findings, Del Boca, Pasqua and Pronzato (2008) confirm that the availability of childcare has a positive effect on probability of working for women. At the same time, family allowances tend to reduce female participation, while they have a minor positive influence on the probability of having a child. In addition, some studies provide evidence for positive effects of urbanisation on labour supply measured at an aggregate level for the whole economy as well as for different age groups of the population (see *e.g.* Grigoli, Koczan and Topalova, 2018).

Modelling retirement behaviour and disentangling the various factors that play a role in people's retirement preferences has attracted a lot of research interest, not least because of its important policy implications for sustainability

of public finances. The literature has identified the concomitant influence of a range of factors such as characteristics of pension schemes, incentives for early retirement, personal income and wealth, health and disability-related factors, demographic factors, individual preferences and the specific household context of older workers. One well-established empirical finding in retirement studies is that reducing generosity of pension schemes induces longer participation in the labour market of the elderly (Behaghel, Blanchet and Roger, 2014). Blundell, French and Tetlow (2016) present strong arguments for the need to analyse joint retirement in the context of collective models of intra-household retirement decision making, pointing at the same time to the lack of uniformly accepted empirical approach for this particular modelling setting. Similarly, in an analysis of the relationship between individual characteristics and the planned retirement age, Riedel, Hofer and Wogerbauer (2015) document a positive correlation between the timing of retirement for partners, when they are already at the stage of planning withdrawal from the labour market.

Kallestrup-Lamb, Kock and Kristensen (2016) empirically investigate the determinants of retirement for Danish workers in 1990 and 1998, adopting a similar approach in dealing with high-dimensional micro data like one of the modern modelling techniques used in this paper. Applying variants of the lasso and the adaptive lasso techniques to logistic regression based on a comprehensive register-based dataset, the authors reduce the size of their model significantly. They finally obtain statistically significant effect on retirement of variables such as age, several labour market indicators, income, wealth as well as a large number of health-related variables.

Finally, a well-documented finding in many studies is that business cycle conditions represent an important factor in explaining labour force participation decisions. Fallick and Pingle (2006) and Grigoli, Koczan and Topalova (2018), among others, show that the expansionary phase of the business cycle represents a statistically significant determinant with a positive influence on the dynamics of the labour force participation rate.

The literature review has shown that a number of personal, household, structural, institutional and cyclical factors intertwine in a complex and dynamic way to shape a person's decision for participating in the labour market against the alternative of being outside of it, with this decision taken in the life-cycle context of each individual. The set of modelling techniques applied in the literature most often includes ordinary least squares regressions, probit and logit models, bivariate probit models and instrumental- variable procedures (general equilibrium models with micro-foundations as well as other structural models are left beyond the scope of this paper).

The potentially very big set of determinants of labour force participation raises challenges for econometric modelling, especially in its classic or more traditional form of econometric models such as those mentioned above. More specifically, drawing meaningful conclusions about the factors driving participation requires empirical approaches that can deal convincingly with important issues such as a large number of potential predictors, possible omitted variables as well as non-linear relationships. The literature review has revealed that a new promising approach dealing with large sets of micro data is the use of regularisation methods (such as variants of the lasso and the adaptive lasso estimators) applied to logistic regressions. A brand new avenue of methodological developments is also the wide and fast spread of various machine learning techniques, which to the best of our knowledge have not yet been used in labour supply literature but have recently been applied in many other economic analyses. With the aim to capture a large initial number of individual and household characteristics as potential drivers of labour force participation in Bulgaria, in this paper we employ two groups of modelling techniques: 1) adaptive lasso and adaptive group lasso which are applied to logistic regression; 2) models based on methods from the machine learning literature. An important advantage of machine learning techniques is that they not only allow encompassing a big number of initial explanatory variables, but can also capture non-linearities. The modelling part of the paper is preceded by a descriptive analysis of the LFS data which is guided by the main findings summarised in the literature review.

3. LFS Micro Data for Bulgaria: Stylised Features

3.1. LFS micro data for Bulgaria: Key Features and Stylised Facts

Our analysis is based on the yearly cross-sectional micro data from the EU's Labour Force Survey (LFS)[1]. The variables in the survey could be divided in three main categories: 1) labour status and employment characteristics, 2) individual characteristics, and 3) household characteristics. The first set of variables includes detailed information about the labour status, hours worked, type of occupation and economic sector classification of the primary and second job, employment search methods, training on the job. Importantly, details on the previous year's labour status and employment are also collected for the

[1] The EU-LFS is the primary labour market data source across EU countries, it is compiled by the national statistical agencies at quarterly and annual frequencies, and processed by Eurostat to ensure comparability across countries. The scope of the EU-LFS and the underlying data quality and comparability standards are laid out in EU legislation. An overview of applicable legal basis with a focus on the use of the anonymised EU-LFS micro data for research purposes can be found in Mack, Lengerer and Dickhaut, 2016. For a detailed description of the full set of data please refer to Eurostat, 2021.

purpose of analysing labour transition outcomes in a short-term perspective. The second set of variables includes various individual characteristics, namely gender, age, country of birth, marital status, level and field of education, region and degree of urbanisation. Furthermore, the LFS contains rich household information, such as household size and composition, number of children and their age, number of inactive or unemployed persons in the household, *etc.*

One very important consideration to note is that – as agreed with Member States – the anonymised EU-LFS micro data does not yet allow tracking people across waves. As the household numbers assigned to individual entries are randomised, it is not possible to trace how the characteristics of a certain individual have changed over time. To this end, at the current juncture the data set cannot be used to assess transition probabilities from one employment status to the other. This caveat of the data set might be revisited going forward – in fact Eurostat has recently launched analyses of the feasibility of constructing longitudinal micro datasets, also in view of their usefulness for individual countries (Eurostat, 2021).

Anonymised EU-LFS micro data for Bulgaria is available for the period 2000–2019[2]. The pooled dataset for this period contains just over 1 million entries.[3] In line with EU-LFS definitions[4], the labour status variable (ILOSTAT) is available for all household members at or above 15 years of age. Those observations represent 89.4 per cent of all entries in the pooled dataset or

[2] The national Labour force survey implements the full set of compulsory variables, according to EU Regulations (https://www.nsi.bg/en/content/12475/metadata/labour-force-survey-annual-data). Details on further aggregations/suppressions by country and year are available in Eurostat, 2021. The following descriptive analysis is based on the full set of micro data available for Bulgaria. Table 4 lists the variables we have included in the empirical analysis – this is after re-coding some variables to ensure consistency over time as explained below.

[3] Entries is used further in the text interchangeably with observations which identify an individual at a specific time point. The EU-LFS is designed as a household survey and the individual entries of the dataset correspond to the individual members of the households, which are interviewed either directly or *via* proxy. The number of respondents (*e.g.* household members directly replying to the survey) for the period 2000–2019 is 613.5 thousand. Therefore, the number of entries/observations in the LFS database is not equal to number of interviewed people at a certain point of time and even less so across time. As already mentioned, the same individuals can be interviewed in different waves of the survey. Also, while for some countries annual surveys may include multiple observations for an individual for different quarters, this does not seem to be the case for the Bulgarian LFS.

[4] "Employment covers persons aged 15 years and over, living in private households, who during the reference week performed work even for just one hour, for pay, profit or family gain, or were not at work but had a job or business from which they were temporarily absent, for example because of illness, holidays, industrial dispute or education and training. Unemployment covers persons aged 15–74 who were not employed during the reference week, were currently available for work and had either been actively seeking work in the past four weeks or had already found a job starting within the next three months". For more information see https://ec.europa.eu/eurostat/statistics-explained/index.php/EU_labour_force_survey.

around 943 thousand unique individual entries (*e.g.* observations which identify an individual at a specific time point).

The dependent variable in the current analysis is *active*, a binary variable derived from the labour status variable. The share of entries with an active labour status amounts to 48.8 per cent (see Figure 2 in Appendix B as regards the annual distribution of the dependent variable). The latter graph also visualises the substantial differences in the number of yearly observations, with entries for 2005, 2006, 2007 being twice as many as compared to other years. At such aggregated level, the dynamics of yearly activity rates (both in terms of unweighted averages or weighted by the special yearly weighting factor[5]) derived from the micro data is very similar to the activity rates from the LFS database for both 15–64 or 15+ age groups (Figure 3 in Appendix B). Namely, it points a gradual increase of activity rates with only a temporary reversal of the trend in the years after the Great Financial Crisis.

Next, we observe the dynamics of activity rates across various subsets as a first and very preliminary step in exploring the relevance of individual, household and employment related factors. The choice of subsets is entirely driven by the factors outlined in the literature review. A closer look at activity rate across time, gender and age groups shows a quite heterogeneous picture both in terms of group-specific levels and dynamics. Figure 4 in Appendix B provides clear evidence on the existence of a hump-shaped relationship with respect to person's age in line with the existent literature. In terms of annual dynamics, the highest increase of the labour participation rates is observed for age groups 55–59 and 60–64 for women, while for men the increase is most evident in the 60–64 age group. This observation could be clearly associated with the gradual increase of the retirement age for men and women.[6]

Figure 5 in Appendix B illustrate how activity rates vary with educational attainment. For both men and women in all age groups these differences are more clear-cut at an aggregate educational level and are fully in line with clearly established empirical importance of educational attainment for positive labour market outcomes. Labour participation also varies across degree of urbanisation, with both men and women in the 15–64 age group living in bigger cities being more likely to be active. Notably, urbanisation seems to be less relevant for younger individuals (*e.g.* in the 15–24 age group), while particularly important for older individuals (*e.g.* above 64 years old) (Figure 6 in Appendix B).

[5] Please refer to the LFS guide for a detailed description of the COEFF variable.

[6] A concise description of the pension reforms undertaken in Bulgaria in the last decade could be found in the EC Ageing reports (see for example European Commission, 2015 and European Commission, 2018).

As regards household characteristics, activity rates also appear to differ across marital status both for men and women in prime age[7] and the relationship is stable across time (Figure 7 in Appendix B). Married men tend to have higher participation rates in line with the established empirical findings in the literature. Surprisingly, married women tend to be more active than singles, while most empirical studies for other countries typically find the opposite. Female participation, however, seems to be negatively related to the number of children in the family and it is very low for women living in households with three or more children. As regards male activity rates, there seem to be no differences in labour outcomes for men with no or up to 2 children. Nevertheless, it is worth mentioning that male activity rates in families with three or more children also seem to be lower and this finding is stable across time (Figure 8 in Appendix B). As could be expected, the age of the youngest child in the family also seems to be relevant for female labour force participation (Figure 9 in Appendix B). Women living in households with children less than 2 years old are considerably less likely to be active, which may be associated with the effects of the statutory maternity leave in the country. Differences in female participation are also observed in families with children below 5 years. On the contrary, women with children in the 6–14 years age group seem to be slightly more active and this relation is also stable across time. Also in line with the literature, male labour participation does not seem to vary much depending on the age of the youngest child in the family, and is actually slightly higher for men living in families with children aged 14 years or less. Another factor that might influence female participation in prime age is the care for elderly members of the household. A closer look at participation rates in families with and without elderly members does not seem to indicate that care for elderly members of the household matters for male labour force participation. For most years covered by the survey female participation in households with elderly dependants was actually higher as these elderly household members could have helped in child-raising. This relation, however, seems to have weakened since 2015 (Figure 10 in Appendix B).

As regards employment characteristics, the labour participation status one year ago seems to be strongly correlated with current activity which is very much in line with the empirical findings on the persistence of labour outcomes[8]. The relation seems to be stable across time and bears equal relevance for men and women (Figure 11 in Appendix B). For both men and women above 40 years of

[7] The relevance of household characteristics is analysed for female and male individuals of typical childbearing age (20–44 years).

[8] We use a subset by the variable WSTAT1Y, which indicates the working status one year before the survey. For the case of Bulgaria, this variable is available only from 2008 onwards.

age, activity rates are close to zero if the person has been inactive one year ago. The probability of transition from inactive to active is higher in the 25–40 age group.

3.2. LFS Micro Data: Constructing the Final Data Subset for Model-Based Analysis

The descriptive analysis presented in the previous section has not only been relevant for drawing stylised facts, but has pencilled certain issues to be addressed before the application of econometric methods for data analysis. Next we provide a short overview of the data cleaning[9] procedures undertaken to construct the final data subset for model-based analysis.

The dependent variable in the current model-based analysis is active, a binary variable derived by recoding the variable on labour status in accordance with the definition of the International Labour Organization, namely *employed + unemployed = active persons*. As already mentioned, the dependent variable is available for all household members at or above 15 years of age or 89.4 per cent of all observations.

Data cleaning with respect to explanatory variables involved the following steps. First, certain data transformations have been performed to handle methodological breaks such as changes in classifications (*e.g.* NACE[10] and occupation (ISCO[11]) classifications, fields and levels of education) or changes in the codification of variables across time (*e.g.* regions). For example, in 2007 the new NUTS2 codes were introduced and while there was no change in the number of regions for Bulgaria, the codes used for these regions changed. Therefore, the re-coding of the variable for earlier years implied that more observations could be kept in the pooled data set. Also as regards the classification of economic activities, professions and education level, re-coding and limited aggregation was undertaken to ensure that the definitions and codes used are largely consistent across time. Since these transformations affected data observations for the years prior to 2007 or 2008, their possible implications for the empirical results, if any, should be very limited – not least because the modelling was eventually carried out on data for the years 2006–2019.

Second, to facilitate the analysis of factors influencing labour participation decisions, similar outcomes have been aggregated in derived variables (*e.g.* number of children, household types, number of children under certain age).

[9] Data cleaning is the process of preparing data for analysis by removing or modifying data that is incomplete, irrelevant, duplicated, or improperly formatted.

[10] Statistical Classification of Economic Activities in the European Community.

[11] International Standard Classification of Occupations.

Third, to avoid an excessive reduction of the number of observations, variables with predominantly missing observations have been dropped from the pooled dataset. Moreover, due to missing observations of key explanatory variables in the 2000–2005 period, the modelling was carried out on data for the years 2006–2019.

Fourth, variables with the property of perfect predictors[12] have been either excluded or re-coded. Most of these variables apply only to active persons by design: sector and field of current employment (for first and second job), work modalities; reasons and duration of unemployment, job-finding techniques, most previous job characteristics. Other variables (*e.g.* number of active, employed and unemployed people in the household) had to be recoded to ensure that they are not perfect predictors by design. Last but not least, a randomly drawn subset of 1500 observations[13] per year has been used for the model-based analysis. This approach was chosen to alleviate the computational burden associated with the estimation of the suite of models used in the analysis, as well as to help reduce the potential bias that can result from an unbalanced sample with differences in the number of observations across years. The final choice of 1500 observations per year was made following preliminary model suite runs to estimate the expected overall solution time and contain it within bounds suitable for iterations and model retraining. The detailed list of variables in the final data subset used in the modelling exercises can be found in Appendix A.

4. Modelling the Determinants of Labour Force Participation

In what follows we present the application of several modelling approaches that aim to uncover the potential driving forces behind labour market participation. The different models we employ are notionally divided into two groups that we refer to as "statistical models" and "machine learning models". These terms, however, may be something of a misnomer in the sense that there is substantial overlap in the tools employed by statisticians and machine learning practitioners. The distinction we draw for the purposes of the present analysis is primarily based on whether a method is chosen because of its theoretical properties that (under certain conditions) uncover facts about the underlying data generating process or, alternatively, the method's predictive power is the

[12] Complete separation in a classification problem, sometimes also referred to as perfect prediction, happens when a predictor variable is sufficient to perfectly classify the values of the outcome variable.

[13] We also draw only one observation per household in order to ensure our sample is i.i.d.

main motivation for using it and understanding the key predictive features follows as a secondary step.

In line with the above principles, subsection 4.1 reports on the application of models that are known to asymptotically select variables that are present in the true data generating process. Subsection 4.2 turns to candidate models that perform well in terms of predictive ability and applies techniques to explain what drives their results. This second group features models that are staples of the machine learning (ML) domain, hence the name we have chosen to employ as a short description.

4.1. Statistical Models

4.1.1. Adaptive Lasso

In this section we provide an overview of the econometric approach used to identify the factors, which can be linked to higher, or respectively lower, labour force participation rate in Bulgaria. We use an adaptive lasso penalised logistic regression to select only the relevant subset of variables from our dataset, and then as a next step, we use a logit regression for the final coefficient estimate. To our knowledge, a similar approach has been applied to a microeconometrics problem only by Kallestrup-Lamb, Kock and Kristensen (2016), where they study the determinants of retirement for Danish workers in 1990 and 1998.

As explained above, the main focus of our paper is to study why economic agents choose to be active or inactive in the labour market, conditional on a set of characteristics we observe about each individual, such as demographics, education, and household type. The logistic regression model, first introduced by McFadden (1973), has become a workhorse model used to study binary choice outcomes. It is solved using Maximum Likelihood Estimation (MLE), where the MLE estimator $\hat{\beta}$ maximises the following objective function:

$$LL(\beta) = \sum_{i=1}^{n} [y_i \ln F({x_i}^T \beta) + (1 - y_i) \ln (1 - F({x_i}^T \beta))] \quad (1)$$

where *LL* is the log-likelihood, y is a binary choice outcome observations vector, x is a vector of explanatory variables, and β is the respective coefficients vector. We use a further augmented version of equation (1) in order to introduce a regularisation penalty on the coefficients β. The literature, particularly in the field of machine learning, in recent years has provided numerous different regularisation or variable selection methods. The approach we choose is the adaptive lasso one (Zou, 2006), which is shown to have good asymptotic properties, and under less strict then other variable selection methods

conditions, to enjoy the oracle property. The oracle property is defined as the property to correctly set the true-zero coefficients of all variables in a set of sparse high-dimensional matrix of explanatory variables, to zero, and at the same time to set the coefficients of the non-zero ones to non-zero values. The MLE's objective becomes minimising the augmented log-likelihood function with the adaptive lasso penalty for a set of p variables:

$$LL^A(\beta) = -\sum_{i=1}^{n}[y_i \ln F({x_i}^T\beta) + (1-y_i)\ln(1-F({x_i}^T\beta))] + \lambda_n \sum_{j=1}^{p} \hat{w}_j|\beta_j|, \quad (2)$$

where λ_n is the standard lasso penalty imposed on the coefficients, and $\hat{w}_j$ are the adaptive lasso weights and are equal to $\hat{w}_j = \frac{1}{|\tilde{\beta}|^\gamma}$. In order to estimate the weights, we first need some initial estimates for $\tilde{\beta}$, which we obtain using a ridge penalised logistic regression. Once the set of the true non-zero parameters has been identified by the adaptive lasso, we again re-estimate the logit model using only the subset of explanatory variables for which the coefficients have not been set to zero. This is a necessary step, as in finite samples the adaptive lasso can introduce bias in the coefficients (Belloni and Chernozhukov, 2013). Our estimation procedure can then be summarised in the following steps[14]:

1. Perform logistic regression with the Ridge penalty to obtain initial $\tilde{\beta}$ estimates.
2. Estimate the logistic regression with the adaptive lasso penalty to identify the true set of non-zero coefficients.
3. Derive the final estimates of the coefficients by solving a logit model using only the subset of variables for which in step 2 the coefficients have been identified as non-zero.

Empirical Estimation

The empirical estimation is performed over a randomly drawn subset of 1500 observations[15] per year, starting 2006. Our initial subset of explanatory variables includes 15 features[16]. Further, the data is split into training and test subsets, the former comprising 70 per cent of the observations. Some continuous variables, such as age, number of inactive people in the household and others,

[14] For the implementation, we rely on the *glmnet* package in R.

[15] We also draw only one observation per household in order to ensure our sample is i.i.d. (independent and identically distributed).

[16] They are SEX, DEGURBA, YEAR, AGE, HATLEV1D, MARSTAT, REGISTER, HHNBOLD, HHCOMP, HHNBEMPL, HHNBUNEM, HHNBINAC, HHNBCHLD_new, HATFIELD_new, REGION_new (for the definition of the variables see Appendix A, Table 4).

are represented instead as categorical variables.[17] The motivation behind this is twofold: it allows the model to capture possible non-linear relationships between the dependent and independent variables, and it ensures all variables are scaled similarly (as dummy variables).[18]

We perform grid-search to find the values of the two hyperparameters λ and γ and choose them based on BIC. In Table 1 we show the estimates of the post-lasso logistic regression.[19] We can see that the adaptive lasso estimator has kept most of the variables in the model. It should be noted, however, that the coefficients, due to the log-odds ratio estimation framework of the logistic regression, do not have a straightforward interpretation. However, we can interpret their sign and relative magnitude. For example, we can see that the coefficients of the YEAR variable show evidence for some pro-cyclical movement of the participation rate. The years of unfavourable macroeconomic environment due to pronounced financial crises such as the euro area sovereign debt crisis, are associated with lower participation rates, whereas years of business cycle upswing (2017 onward) are linked with a positive probability of participating in the labour force. Furthermore, the negative coefficient of SEX_2 (female) indicates that gender plays a role in labour supply, reducing the probability of labour market participation.

Another variable that stands out due to the magnitude of its coefficients is HATLEV1D_L (having low education). We can see that education is pivotal for being part of the labour force, and people with low education status are more likely to be inactive.

[17] The variable age is aggregated into four distinct categories: Age-1 – people aged under 22 years, Age-2 – between 22 and 55 years, Age-3 – between 55 and 65 years, and Age-4 – older people. The grouping is built on the assumption that youths (young people under 22 completing their education or just entering the labour market), people in their prime working age (people aged 22–55), people approaching retirement (people between 55 and 65 years) and older people might show different labour force participation behaviour.

[18] Lasso requires all variables to be either of the same unit or scaled.

[19] The accuracy of the model is 0.85 and the kappa score 0.69.

Table 1: **Adaptive Lasso Estimates**

	Short Description	Coef.	Std. Err.		Short Description	Coef.	Std. Err.
SEX_2	Sex of the respondent	-0.7272***	0.053	HHNBINAC_1.0	Number of inactive adults in the household	-0.6963***	0.074
DEGURBA_2.0	Degree of urbanisation	-0.0814	0.075	HHNBINAC_2.0		-0.5314***	0.134
DEGURBA_3.0		-0.0089	0.061	HHNBINAC_3.0		-0.4218*	0.250
YEAR 2011	Fixed reference year	-0.1189	0.100	HHNBINAC_4.0		-0.2329	0.526
YEAR 2012		-0.2703***	0.100	HHNBCHLD_new_2.0	Number of children in the household	-0.1609	0.100
YEAR_2013		-0.1169	0.099	HHNBCHLD_new_4.0		-0.2188	0.323
YEAR_2017		0.4083***	0.104	HATFIELD_new_1.0	Field of education	0.1303	0.202
YEAR 2018		0.1091	0.105	HATFIELD_new_2.0		0.2022	0.261
YEAR 2019		0.3727***	0.105	HATFIELD_new_3.0		0.3748***	0.130
AGE_2	Age of interviewed person	1.9465***	0.103	HATFIELD_new_6.0		0.2494***	0.084
AGE_3		0.5886***	0.119	HATFIELD_new_8.0		0.3780*	0.225
AGE_4		-1.7773***	0.166	HATFIELD new 999.0		0.2426***	0.089
HATLEV1D_L	Level of education	-2.0193***	0.104	REGION_new_33	Region of household (NUTS2 classification)	0.2518***	0.086
HATLEV1D_M		-0.8545***	0.081	REGION_new_34		0.1509*	0.085
MARSTAT_1	Marital status	-0.1068	0.092	REGION_new_41		0.2081***	0.071
MARSTAT_2		0.3921***	0.088	REGION_new_42		0.3080***	0.075
REGISTER_2.0	Registration at a public employment office	0.8060***	0.206	HHNBEMPL_1.0	Number of employed adults in the household	0.4567***	0.064
REGISTER_4.0		0.299**	0.151	HHNBEMPL_3.0		0.7021***	0.169
REGISTER_9.0		-1.3688***	0.262	HHNBEMPL_4.0		0.5117	0.338
HHNBOLD_1.0	Number of persons aged 65 or older in the household	-0.2749***	0.087	HHNBUNEM_1.0	Number of unemployed adults in the household	0.5213***	0.120
HHNBOLD_2.0		-0.5127***	0.146	HHNBUNEM_2.0		0.1961	0.338
HHCOMP_12.0	Households composition	-0.4786**	0.232				
HHCOMP_20.0		0.1846*	0.096				
HHCOMP_21.0		-0.1660	0.121				
HHCOMP_22.0		-0.4594***	0.135				
HHCOMP_30.0		0.5069***	0.087				
HHCOMP_31.0		0.3523***	0.121				

***p < 0.01, **p < 0.05, *p < 0.1
For the definition of the variables see Appendix A, Table 4.

4.1.2. Adaptive Group Lasso

The variable selection the adaptive lasso performs is done on an individual-parameter basis. However, when working with categorical variables, which are then transformed into dummy variables corresponding to each category, it is intuitive to think that the selection should be performed on a group level for all same-category dummies. This is handled by the group lasso, first introduced for logistic regression by Meier, Geer and Buhlmann (2008). The penalty term in the objective lasso function, shown in equation (2), is modified in order to penalise a whole group of variables, rather than individual ones. This ensures that the lasso method will either set the coefficients of all variables in the group

to zero, or none of them. The parameter vector $\hat{\beta}$ then minimises the following convex objective function:

$$S(\beta) = -LL(\beta) + \lambda \sum_{g=1}^{G} s(p_g)||\beta_g||_2 \tag{3}$$

where *LL(...)* is the log-likelihood function of the logistic regression, $g = 1...G$ are the number of groups and λ is the penalty. The penalty term, defined as the sum of the Euclidean norms of the group-specific coefficient vectors, ensures that the model performs variable selection at the group level. In practice, it is a combination of the lasso l_1 and the ridge l_2 penalties. If all groups had one variable each, the penalty would be the same as the standard lasso penalty. If, however, there was only one group with multiple variables included in it, the penalty would reduce to the ridge one. This intermediary between the two ensures that sparsity is introduced at the group level, but however no sparsity is introduced within the group itself. The function $s(p_g)$ is used to account for the dimensionality of β_g, or in other words to rescale the penalty, conditional on the number of variables in each group. It is common in the literature to set $s(p_g) = p_g^{0,5}$ (Yuan and Lin, 2006). Meier, Geer and Buhlmann (2008) further elaborate on the asymptotic properties of the logistic group lasso and show that, under some regularity and sparseness conditions, $\hat{\beta}$ is a globally consistent estimator of β. As a method of choosing the value of the hyperparameter λ, they suggest using $\ln(G)$.

In order to ensure both parameter and variable selection consistency (Oracle property), we further alter the penalty term in equation (4). We use the framework introduced by Wang and Leng (2008), who present an implementation of the adaptive group lasso weights in the linear case. We built on that by applying the adaptive weights to a logistic group lasso. In our specification, the objective function is changed the following way:

$$S^A(\beta) = -LL(\beta) + \sum_{g=1}^{G} s(p_g)\lambda_g||\beta_g||_2 \tag{4}$$

where $\lambda_g = \lambda||\tilde{\beta}_g||^{-\gamma}$. The adaptive group lasso modification allows us to impose a different penalty per group λ_g, based on initial consistent coefficients β_g. Intuitively, this means that if some of the coefficients in the group variables are close to zero, the group will be given a larger penalty, and *vice versa*. Wang and Leng (2008) show that in the linear case this yields an estimator with the Oracle property.

For the initial $\tilde{\beta}_g$ we consider the estimates of the non-penalised logistic regression, and perform grid-search for the values of λ and γ. We choose the values minimising the BIC, where:

$$BIC = \overbrace{-2LL}^{\text{Log-Likelihood}} + \overbrace{\ln(N)df}^{\text{Parameters Penalty}} \tag{5}$$

Note that for the BIC estimation we define the degrees of freedom[20] similarly to Yuan and Lin (2006):

$$df = \sum_{g=1}^{G} I\left\{\hat{\beta}_g > 0\right\} + \sum_{g=1}^{G} \frac{||\hat{\beta}_g||}{||\tilde{\beta}_g||}(p_g - 1) \tag{6}$$

Empirical Estimation

Estimation is performed in a similar manner to one the described for the adaptive lasso model.

Figure 1 shows the group adaptive lasso results. The *y*-axis shows each explanatory variable's estimated coefficient, and on the *x*-axis we see the variables, represented by their arbitrary ordering when entering the model (variables from the same group enter the model consecutively and therefore have consecutive indices on the *x*-axis). It is straightforward to see that the estimator has set the coefficients of entire groups (rather than individual variables) to zero. The variables that are at the end selected by the model are four (out of 15): age, level of education, number of employed and inactive household members.

[20] We use model degrees of freedom, rather than residuals degrees of freedom.

Figure 1: **Group Adaptive Lasso Estimation**

As a last step in the estimation, we obtain the coefficients using post-lasso logit[21] with only the variables selected by the group adaptive lasso. The results are presented in Table 2. The model chooses only four variables as the ones with the most explanatory power: age, level of education, number of employed household members, and number of inactive household members. In terms of age, we can see evidence for the expected non-linearity in how labour force participation depends on the age of the individual. Being in the age group of 22 to 55 years has a strong positive correlation with participating in the labour force. As age increases, the positive co-movement diminishes, and the correlation turns negative for people aged 65 years and older. Low level of education also has strong negative correlation with labour force participation. The labour decisions of the rest of the household members also are related to the individual's decision, but to a lesser degree, compared to age and level of education.

[21] The accuracy of the model is 0.84 and the kappa score 0.65.

Table 2: **Post-lasso Estimation**

	Short Description	Coef.	Std. Err.
AGE_2	Age of interviewed person	2.4217***	0.055
AGE_3		1.1449***	0.055
AGE_4		-2.0361***	0.070
HATLEV1D–L	Level of education	-2.0203***	0.060
HATLEV1D_M		-0.8466***	0.053
HHNBEMPL_1.0	Number of employed adults in the household	0.6600***	0.048
HHNBEMPL_2.0		0.4252***	0.070
HHNBEMPL_3.0		1.2345***	0.128
HHNBEMPL_4.0		0.5594*	0.286
HHNBINAC_1.0	Number of inactive adults in the household	-0.5688***	0.047
HHNBINAC_2.0		-0.7370***	0.091
HHNBINAC_3.0		-0.4657**	0.213
HHNBINAC_4.0		-0.2516	0.427

*** p < 0.01, ** p < 0.05, * p < 0.1

Note: For the definition of the variables see Appendix A, Table 4.

4.2. Machine Learning Models

4.2.1. General Considerations

ML techniques have earned a reputation for their high predictive ability.[22] This makes them strong candidates for consideration in situations where forecast accuracy is the primary goal of modelling. The variety of approaches embedded in ML methods – nonlinear techniques, resampling, combination of the results of individual predictors, *etc.* – provide a sufficiently broad and flexible basis for gains in prediction accuracy.

While ML methods may offer added value in a pure forecasting context, a well-recognised downside is that the embedded complexity that assists in achieving accurate predictions also makes the models and their results harder to understand and interpret. The standard solution has been to relegate ML models to applications where interpretability takes a back seat to forecasting capability, and resort to more traditional statistical methods when interpretation is of primary importance.

With ML models finding an ever expanding scope of application over the past decade, interest in their interpretability and "explainability" has peaked,

[22] An often-cited example is the prevalence of machine learning models as winners in forecasting competitions (*e.g.* Kaggle). Results such as the ones reported in Makridakis, Spiliotis and Assimakopoulos (2020) partially support this claim, although they paint a much more nuanced picture.

spawning a new strand of research in the field (Belle and Papantonis, 2020). The motivation for this goes beyond a purely academic need for improved understanding and crosses over to issues such as ethics, potential bias and discrimination. In any case, a number of tools have been developed recently to assist in understanding the functioning of ML models and in explaining their output (Molnar (2019) and Biecek and Burzykowski (2021) are contemporary works in textbook format that offer an introduction to the field).

The above considerations suggest that ML methods can be applied to model labour force participation in an attempt to blend high predictive performance and interpretability of the results. This approach is potentially useful in that it can exploit the strengths of ML techniques and uncover associations between variables in the data that may be missed by conventional econometric methods. Specifically, the ML methods employed may help to capture nonlinearities and complex interactions between the different variables. At the same time, the subsequent application of model interpretation tools can help open the black boxes and extract explanations about the model behaviour and the results obtained.

In what follows, this strategy is implemented by selecting several popular ML models and training them on a subset of the LFS data. The results of the training step are then fed into a set of model interpretation procedures to obtain explanations of the relative importance of the different variables, study potential non-linearities with respect to the variables that manifest a sufficient degree of importance for model results and, finally, look at the possibility of explaining the driving forces behind predictions for particular observations of interest.

We first describe the main characteristics of the different methods employed, stressing the intuition and providing references to more detailed and technically complete presentations. Then, we elaborate on the use of the tools for modelling labour force participation and discuss the results.

4.2.2. Overview of Selected ML Models and Interpretation Techniques

ML models Our selection of models comprises four classification models: elastic net (Enet), support vector machine (SVM), random forest (RF) and K-nearest neighbours (KNN). We did not aim to be exhaustive and include as many models as possible. Rather, the goal was to include representatives of several different modelling approaches and see what each of them can contribute to explaining the determinants of labour force participation.

The elastic net model is a representative of the class of penalisation methods. The approach was originally proposed by Zou and Hastie (2005). The main idea of the approach is that the incorporation of variables in the model, as well

as their magnitude, can be controlled by modifying the objective function in a logistic regression to include appropriate penalty terms. These terms reward sparsity, thus effectively implementing a variable selection procedure, and shrink the coefficients of correlated variables. Tuning parameters control the weight of each of these effects in the estimation (see Hastie, Tibshirani and Friedman (2009), Chapter 18.4, for more information on the model).

SVMs provide an approach to classification that can be thought of as a generalisation of the idea of separating hyperplanes. An early version of the method was proposed in Cortes and Vapnik (1995), though its origins trace back to research from the 1960s. Intuitively, a SVM constructs a (generally non-linear) boundary that tries to separate the different classes. As clean separation is possible only in stylized problems, the approach aims to minimise classification errors while controlling for overfitting. Different implementations exist and the models in this class are tunable to control the tradeoff between robustness and model fit. Further information on the method can be found in James et al. (2013), Chapter 9, or Murphy (2012), Chapter 14.

Random forests are a class of ensemble models (*i.e.* models combining individual classifiers). Random forests were originally proposed in Tin Kam Ho (1995) and later extended by Breiman (2001). The main idea is to fit a number of decision tree models on bootstrapped training samples and combine their predictions according to some rule to arrive at an aggregated prediction. The procedure for fitting the decision trees is modified to include random constraints on the subsets of variables considered. This approach help to reduce the correlation between the individual tree predictions, which improves forecast accuracy. Random forests are frequently among the top performers in terms of predictive ability. Details on the method can be found in Hastie, Tibshirani and Friedman (2009), Chapter 15, or James et al. (2013), Chapter 8.

KNN models are strongly data-dependent in the sense that they always work with a particular dataset, rather than extracting estimates that, once obtained, are decoupled from the data. KNN models were suggested in Fix and Hodges (1951). The implement the approach of constructing a neighbourhood of a particular observation by means of an appropriate distance function. The neighbourhood will contain a predefined number of points (the K parameter) that are closest to the observation in question. A classification decision is then taken based on the characteristics of the points in the respective neighbourhood. More information can be found in Alpaydin (2010), Chapter 8.

Model interpretation Model interpretation methods can be classified in different groups depending on whether:

- they operate at the dataset or individual observation level (global or local methods);
- they exploit inherent interpretability in the structure of the model (*e.g.* as in linear regression models) or are model-agnostic.

In what follows, we work exclusively with model-agnostic methods. The advantages of this approach are that, first, we can analyse both interpretable and black-box methods using it and, second, our methodological framework can easily be extended to incorporate additional models and they will be treated symmetrically to the ones already included in the ML model suite. Of the methods presented below, permutation feature importances and partial dependence plots are global, *i.e.* they characterise the models as a whole, while local interpretable model-agnostic explanations (LIME) and Shapley values are local methods which aim to explain individual predictions.

Permutation feature (variable) importances exploit the idea that model performance will degrade if information from variables that matter is removed. Model performance in a ML context is usually defined in terms of prediction accuracy and one way of removing information about a variable is to permute randomly its values. The extent to which the prediction error of the model increases when using the permuted variable is a measure of the importance of this variable. The way to measure the prediction error can vary, depending on the context. Here we use the model classification error as a metric for predictive accuracy. More details on permutation feature importances can be found in Molnar (2019), Chapter 5.5, or Biecek and Burzykowski (2021), Chapter 16.

Partial dependence plots provide a way to measure the effect of varying one of the model variables on the model predictions. In order to do that, one has to come up with a way to treat the other variables in the model. The construction of partial dependence plots handles this by taking the average of the predictions for the values of the other variables in the dataset. Thus, partial dependence plots measure the average marginal effect of a variable on the prediction. This is intuitive and straightforward to compute but may distort the dependence pattern in the case of correlated independent variables. A more extensive discussion is available in Molnar (2019), Chapter 5.1, or Biecek and Burzykowski (2021), Chapter 17.

LIME methods explain the prediction of a complex model by estimating a local approximation of the predictions around a particular observation by means of an interpretable model such as a linear regression or a decision tree. In essence, the black box model to be interpreted is used as a data generator to provide estimation samples for the interpretable model. These samples are weighted appropriately to account for the proximity to the point of interest. The

output of LIME is a subset of variables and their respective contributions to a particular prediction. Chapter 5.9 in Molnar (2019) contains further details on the method.

Yet another approach to explaining an individual prediction is given by the Shapley value. The Shapley value is a concept from cooperative game theory that proposes an equitable distribution of a payoff among the different players. In a model interpretation context this can be re-implemented as a game that "distributes" a prediction among the different explanatory variables according to their contributions to the prediction. See Molnar (2019), 5.9 and 5.10, and Biecek and Burzykowski (2021), Chapter 8, for a presentation of different local explanation methods based on the Shapley value.

4.2.3. Determinants of LFP: Machine Learning Techniques

The sample we used to train the ML models for the decision to participate on the labour market comprises the period 2006–2019. Due to structural changes in the data, the period 2000–2005 was excluded from the modelling exercise. Since the ML model training is computationally intensive for large data sets, we downsampled the data for each year to include 1500 observations.[23] The particular annual sample size was chosen after several experiments with different sample sizes in which model solution times were explored. The final sample size choice was judged to provide a reasonable compromise between data coverage and manageable total solution time for the suite of models. In addition to providing a more manageable overall sample size, this approach serves to alleviate problems with unbalancedness arising from differences in sample sizes over the different years.

As the LFS data contain a number of variables that are perfect predictors for the labour market status of a person, we chose the maximal subset of variables that can serve to explain labour market participation without being perfectly correlated with it. The gist of our approach is to start with as many predictors as possible in order to minimise the risk of omitting relevant information and let the ML models determine which predictors are important drivers of participation. We ended up with a set of 16 explanatory variables.[24]

Model training was implemented in R using the infrastructure provided by the *caret* package. The data was split into training and testing sets with 70 per cent

[23] The *createDataPartition* function from R package *caret* was used in this procedure. The function tries to extract a representative sample from the full dataset.

[24] In addition to the dependent variable ILOSTAT, the explanatory variables are SEX, DEGURBA, YEAR, AGE, HATLEV1D, MARSTAT, REGISTER, HHNBOLD, HHCOMP, HHNBEMPL, HHNBUNEM, HHNBINAC, HHNBCHLD_new, HATFIELD_new and REGION_new. Appendix A, Table 4 contains the precise definitions of the different variables.

of the observations used for ML model training and the remainder used for out-of-sample evaluation. All models were tuned through repeated cross-validation with 10 folds and 5 repeats. The data were centred and scaled as required by the respective methods. The specific *caret* methods used for the training procedure were as follows:

- Enet – *glmnet*
- SVM – *svmRadial*
- RF – *rf*
- KNN – *knn.*

Table 3: **Variables with High Permutation Feature Importance for the Suite of Models**

Variable	Short description	ENet	SVM	RF	KNN
AGE	Age of interviewed person	*	*	*	*
HHNBINAC	Number of inactive adults	*	*		
REGISTER	Registration at a PE office	*	*		
HATLEV1D	Level of education	*	*	*	
HATFIELD_new	Field of education		*	*	*
REGION_new	Region of household			*	
HHNBOLD	Number of persons aged > 65			*	
HHCOMP	Households composition				*
HHNBEMPL	Number of employed adults	*		*	
SEX	Sex of the respondent	*			
YEAR	Fixed reference year		*	*	*

Following model training, the out-of-sample predictive performance of the suite was checked on the test data. The elastic net, SVM and random forest models achieved prediction accuracy of about 0.86 and kappa of approximately 0.72. The performance of the KNN model was slightly worse, with prediction accuracy of 0.84 and kappa of 0.69. Based on these results, we deemed all four models in the suite as possessing sufficiently high predictive power and proceeded to apply model interpretation techniques to the tuned models.

Figures 12–15 in Appendix C present the permutation feature importances for the models in the ML suite. We set thresholds for importances for each model based on visual inspection of the plots to identify kinks in the results. Table 3 summarises our findings and identifies the points of agreement between the models with respect to variable importance.

The results reported in Table 3 indicate a consensus on the importance of the age of the person as explanatory factor for their labour market status. The educational level of the person and their field of education also play an important role, as

they are picked up by three of the four models. The year of the observation is also identified as important by three models, underscoring the contribution of time-specific factors such as the cyclical state of the economy. There is some evidence in favour of the importance of the number of inactive persons in the household, registration at a public employment office and the number of employed persons in the household. The results suggest that there is little support for the number of old persons in the household, the region, the sex of the person and the household composition as determinants of labour market status.

Further information on the nature of the relationships picked up by the models can be obtained from the partial dependence plots for the important variables in the different models, as presented in Figures 16–19 in Appendix C. An important result that corresponds to stylised facts from the literature is the hump-shaped dependence of participation on the age of the person. Predictably, the higher the education level of the person, the higher the probability of participation is. In terms of the field of education, there is evidence that Information and Communication Technology specialists are associated with a higher probability of being active, and some evidence on the importance of having a specialisation in health and welfare, as well as in the broad groups *Humanities, languages and arts* and *Natural sciences, mathematics and statistics*. The partial dependence plots for the year of the survey indicate that the probability of being active has increased somewhat towards the end of the sample.

Figures 20 and 21 provide two examples of local interpretations for a specific data instance, LIME for the KNN model and the Shapley value for the random forest model. In the case of the LIME explainer a small subset of the variables was identified as important for the prediction. The figure shows, first, that having an old person in the household and having marital status 0 (Widowed, divorced or legally separated) negatively affects the probability of labour market participation and, second, in this particular case the number of old persons has a larger negative contribution to the probability of participation. Similar interpretations can be obtained from the Shapley value explainer, which does not restrict the number of variables used in the explanation and therefore identifies the age of the person as the primary factor affecting the probability of participation.

4.3. Discussion of the Results

Overall, both the statistical and the ML approaches confirm the role of a person's age as a determinant of labour market participation. Our results broadly reproduce the stylised fact that the probability of participation is hump-shaped when viewed as a function of age. The positive association between the level of education and the probability of labour market participation is also a

point of consensus among the models. Other aspects of a person's educational profile, specifically their field of education, appear to play a role as well, though these seem less prominent than the level of education.

Additionally, there is strong evidence that characteristics such as the number of household members that are employed or inactive also correlate with labour market participation. A tentative interpretation of this finding may be that the overall household situation matters and it may encourage or deter from actively participating in the labour market. However, further analysis is needed to confirm such a causal interpretation and this is left as a possible direction for future work.

We have also found some evidence supporting the importance of time-specific factors for the probability of labour market participation. These in part reflect the cyclical position of the economy but also subsume other impacts specific to the respective period. There are also some empirical indications that gender plays a differentiating role for labour market status, as results show that women have lower labour force participation. However, the degree of agreement among the different models is much smaller on that finding.

5. Concluding Remarks

Addressing an important policy issue about labour supply determinants in Bulgaria, this paper has analysed the relevance of different socioeconomic factors for the decision of individuals to participate in the labour market. Theoretical and empirical studies on labour force participation behaviour point to a wide range of potential determinants such as personal and household characteristics, cyclical, structural as well as institutional factors. A common finding in many studies is the importance of studying labour supply decisions within the wider context of intra-household decision-making.

In this paper, we use an anonymised micro dataset from the Labour Force Survey for Bulgaria over the period from 2000 to 2019. A key benefit of the data is that it allows a very detailed level of focus on the driving forces behind labour market participation, particularly those related to individual and household characteristics. The descriptive analysis in the paper has shown the potential importance of age, educational attainment and urbanisation for economic activity. Furthermore, female participation appeared to be negatively related to the number of children in the family and particularly low for women living in households with three or more children. In the latter case, we also observe relatively low male economic activity. Women with children less than 2 years old are also considerably less likely to be active, which may be associated with the effects of the statutory maternity leave in the country. The presence of elderly people in the family does not seem to be a relevant factor for male labour force

participation. While for most years covered by the survey female participation in households with elderly members was actually higher, this relation seems to have weakened since 2015. Finally, the labour force participation status of a person one year ago appeared to have a strong correlation with present activity, in line with empirical findings on the persistence of labour outcomes.

We explored the relevance of a broad set of potential factors for the labour force participation decision based on the extensive information available in the Labour Force Survey micro data set. In order to be able to address convincingly important statistical issues such as a large potential number of predictors, possible omitted variables as well as non-linear relationships, we relied on a set of complex modelling techniques. The techniques employed for the purposes of the analysis included adaptive lasso and adaptive group lasso from the statistical domain and various methods from the machine learning literature. Our main results corroborate some of the hypotheses from the descriptive analysis. A common result from the modelling approaches is the hump-shaped relationship of labour force participation with respect to person's age – a stylized fact from the literature, which we also noted in our descriptive section. A robust finding from all models is the clear positive association between the level of a person's education and the probability of being economically active. Concurrently, there is some evidence for the additional role played by the field of education, although not all models have confirmed this. Another robust finding that we obtain is the importance of specific household characteristics such as the number of household members that are employed or inactive. We find that the number of employed household members tends to be positively linked with participation probability, while that of inactive people lowers that probability. We also find empirical evidence that the upward phase of the business cycle is positively linked with the probability of being economically actively. In terms of economic policy implications, based on all models employed we may convincingly conclude that education is a key contributing factor for being part of the labour force.

We identify the following areas for future research work. While our modelling techniques have enabled us to work with a large set of potential predictors, we have not addressed in a strict theoretical way the casual link between these predictors and labour force participation probability. For example, studying in more detail the causal association between household characteristics and economic activity appears a promising area for analysis. Future research could also focus on overcoming potential endogeneity issues with other modelling techniques such as instrumental variables. Furthermore, double machine learning techniques could be used for obtaining unbiased estimates of specific parameters of interest.

References

Alpaydin, E. 2010. *Introduction to Machine Learning*. 2nd. The MIT Press. ISBN: 026201243X.

Angrist, J., and Evans. 1998. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *The American Economic Review* 88 (3): 450477. ISSN: 00028282. https://www.jstor.org/stable/116844.

Azcona, G., A. Bhatt, W. Cole, R. Gammarano and S. Kapso. 2020. *The Impact of Marriage and Children on Labour Market Participation. Spotlight* on Goal 8. ILO – UN Women. https://data.unwomen.org/sites/default/files/inline-files/Spotlight-goal8-spread.pdf.

Bachmann, R., D. Baumgarten, H. Kroger, S. Schaffner, M. Vorell and M. Fertig. 2010. *Study on various aspects of labour market performance using micro data from the European Union Labour Force Survey*. RWI – Leibniz-Institut fur Wirtschaftsforschung. https://econpapers.repec.org/RePEc:zbw:rwipro:69936.

Becker, G. S. 1965. "A Theory of the Allocation of Time." *The Economic Journal* 75 (299): 493–517. ISSN: 00130133, 14680297. https://www.jstor.org/stable/2228949.

Becker, G. S. 1974. "A Theory of Social Interactions." *Journal of Political Economy* 82 (6): 1063–1093. ISSN: 00223808, 1537534X. https://www.jstor.org/stable/1830662.

Behaghel, L., D. Blanchet and M. Roger. 2014. *Retirement, Early Retirement and Disability: Explaining Labor Force Participation after 55 in France*. Working Paper, Working Paper Series 20030. National Bureau of Economic Research, April. https://www.nber.org/papers/w20030.

Belle, V., and I. Papantonis. 2020. *Principles and Practice of Explainable Machine Learning*. arXiv 2009.11698, September. arXiv: 2009.11698 [cs.LG].

Belloni, A., and V. Chernozhukov. 2013. "Least squares after model selection in high- dimensional sparse models." *Bernoulli* 19, no. 2 (May): 521–547. doi:10.3150/11- BEJ410. http://dx.doi.org/10.3150/11-BEJ410.

Biecek, P., and T. Burzykowski. 2021. *Explanatory Model Analysis*. Chapman / Hall/CRC, New York. ISBN: 9780367135591. https://pbiecek.github.io/ema/.

Blundell, R., E. French and G. Tetlow. 2016. "Chapter 8 – Retirement Incentives and Labor Supply," edited by J. Piggott and A. Woodland, 1:457–566. Handbook of the Economics of Population Aging. North-Holland. http://www.sciencedirect.com/science/article/pii/S2212007616300190.

Blundell, R. 1995. "The Impact of Taxation on Labour Force Participation and Labour Supply," no. 8. https://www.oecd-ilibrary.org/content/paper/576638686128.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45, no. 1 (October): 5–32. ISSN: 1573-0565. https://doi.org/10.1023/A:1010933404324.

Cipollone, A., E. Patacchini and G. Vallanti. 2013. *Women Labor Market Participation in Europe: Novel Evidence on Trends and Shaping Factors*. IZA Discussion Papers 7710. Institute of Labor Economics (IZA), October. https://ideas.repec.org/p/iza/izadps/dp7710.html.

Clark, K. B., and L. H. Summers. 1982. *Labor Force Participation: Timing and Persistence*. Working Paper, Working Paper Series 977. National Bureau of Economic Research, September. http://www.nber.org/papers/w0977.

Cortes, C., and V. Vapnik. 1995. "Support-Vector Networks." *Machine Learning* (USA) 20, no. 3 (September): 273–297. ISSN: 0885-6125. https://doi.org/10.1023/A:1022627411411.

Del Boca, D., S. Pasqua and C. Pronzato. 2008. "Motherhood and market work decisions in institutional context: a European perspective." *Oxford Economic Papers* 61, no. suppl_1 (December): i147-i171. ISSN: 0030-7653. https://doi.org/10.1093/oep/gpn046.

European Commission. 2015. *Ageing Report, 2015*.

European Commission. 2018. *Ageing Report, 2018*.

Eurostat, N. 2021. "EU Labour Force Survey database user guide." doi: http://dx.doi.org/https://ec.europa.eu/eurostat/documents/1978984/6037342/EULFS-Database-UserGuide.pdf.

Fallick, B., and J. Pingle. 2006. *A cohort-based model of labor force participation*. Finance and Economics Discussion Series 2007-09. Board of Governors of the Federal Reserve System (U.S.) https://econpapers.repec.org/RePEc:fip:fedgfe:2007-09.

Fix, E., and J. L. Hodges. 1951. *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*. Technical Report 4. USAF School of Aviation Medicine.

Genre, V., R. Gomez-Salvador and A. Lamo. 2005. *European women: Why do(n't) they work?* Working Paper Series 454. European Central Bank, March. https://ideas.repec.org/p/ecb/ecbwps/2005454.html.

Grigoli, F., Z. Koczan and P. Topalova. 2018. *Drivers of Labor Force Participation in Advanced Economies: Macro and Micro Evidence*. IMF Working Papers 18/150. International Monetary Fund. doi: http://dx.doi.org/10.5089/9781484361528.001.

Hastie, T., R. Tibshirani and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York. ISBN: 9780387848587.

James, G., D. Witten, T. Hastie and R. Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications* in R. Springer Publishing Company, Incorporated. ISBN: 1461471370.

Kallestrup-Lamb, M., A. B. Kock and J. T. Kristensen. 2016. "Lassoing the Determinants of Retirement." *Econometric Reviews* 35 (8–10): 1522–1561. eprint: https://www.tandfonline.com/doi/pdf/10.1080/07474938.2015.1092803.

Killingsworth, M. R., and J. J. Heckman. 1986. "Chapter 2 – Female labor supply: A survey," 1:103-204. Handbook of Labor Economics. Elsevier. http://www.sciencedirect.com/science/article/pii/S1573446386010052.

Mack, A., A. Lengerer and T. Dickhaut. 2016. *Anonymized EU-LFS Microdata for Research: Background, Resources, and Introduction to Data Handling*. 2016/15:39. GESIS Papers. Koln: GESIS – Leibniz-Institut fur Sozialwissenschaften.

Makridakis, S., E. Spiliotis and V. Assimakopoulos. 2020. "The M4 Competition: 100,000 time series and 61 forecasting methods." *International Journal of Forecasting* 36 (1): 54-74. ISSN: 0169-2070. http://www.sciencedirect.com/science/article/pii/S0169207019301128.

McFadden, D. 1973. "Conditional Logit Analysis of Qualitative Choice Behaviour." In *Frontiers in Econometrics*, edited by P. Zarembka, 105–142. New York, NY, USA: Academic Press New York.

Meier, L., S. van de Geer and P. Buhlmann. 2008. "The group LASSO for logistic regression." *Journal of the Royal Statistical Society Series B* 70 (February): 53-71. doi: http://dx.doi.org/10.1111/j.1467-9868.2007.00627.x.

Mincer, J. 1962. "Labor Force Participation of Married Women: A Study of Labor Supply." In *Aspects of Labor Economics*, 63–105. Princeton University Press. http://www.nber.org/chapters/c0603.

Molnar, C. 2019. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/.

Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press. ISBN: 0262018020.

Pencavel, J. 1986. "Chapter 1 – Labor supply of men: A survey," 1:3–102. Handbook of Labor Economics. Elsevier. http://www.sciencedirect.com/science/article/pii/S1573446386010040.

Riedel, M., H. Hofer and B. Wogerbauer. 2015. "Determinants for the transition from work into retirement in Europe." *IZA Journal of European Labor Studies* 4, no. 1 (February). ISSN: 2193-9012. https://doi.org/10.1186/s40174-014-0027-5.

Strauch, R., R. Gomez-Salvador, M. Ward-Warmedinger, J. Turunen, N. Leiner-Killinger and K. Masuch. 2008. *Labour supply and employment in the euro area countries: developments and challenges*. Occasional Paper Series 87. European Central Bank, June. https://ideas.repec.org/p/ecb/ecbops/200887.html.

Tin Kam Ho. 1995. "Random decision forests." In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, 278–282 vol.1. doi: http://dx.doi.org/10.1109/ICDAR.1995.598994.

Vlasblom, J. D., and J. J. Schippers. 2004. "Increases in Female Labour Force Participation in Europe: Similarities and Differences." *European Journal of Population / Revue Europeenne de Demographie* 20 (4): 375–392. ISSN: 01686577, 15729885. http://www.jstor.org/stable/20164280.

Wang, H., and C. Leng. 2008. "A note on adaptive group lasso." *Computational Statistics Data Analysis* 52 (12): 5277–5286. ISSN: 0167-9473. http://www.sciencedirect.com/science/article/pii/S0167947308002582.

Yuan, M., and Y. Lin. 2006. "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society Series B* 68 (1): 49–67. https://econpapers.repec.org/RePEc:bla:jorssb:v:68:y:2006:i:1:p:49-67.

Zou, H. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101 (476): 1418–1429. eprint: https://doi.org/10.1198/016214506000000735.

Zou, H., and T. Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301–320. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2005.00503.x.

A Appendix

Table 4: **Final LFS variables used in the empirical analysis and their abbreviations**

Code	Definition
ID	**Key variable**
ILOSTAT	**Labour status;** 1 – Active 2 – Inactive
SEX	**Sex of the respondent;** 1 – Man 2 – Female
DEGURBA	**Degree of urbanisation;** 1 – Cities (Densely-populated area) 2 – Towns and suburbs (Intermediate density area) 3 – Rural area (Thinly-populated area)
INTWAVE	**Sequence number of the survey wave;** 1–4
YEAR	**Fixed reference year**
AGE	**Age of interviewed person**
HATLEV1D	**Level of education** L – Low M – Medium H – High
MARSTAT	**Marital status** 0 – Widowed, divorced or legally separated 1 – Single 2 – Married
REGISTER	**Registration at a public employment office** 1 – Person is registered at a public employment office and receives benefit or assistance 2 – Person is registered at a public employment office but does not receive benefit or assistance 4 – Person is not registered at a public employment office and does not receive benefit or assistance 9 – Not applicable (person aged less than 15 years or older than 74)
HATYEAR	**Year when highest level of education was successfully completed**
HHNBOLD	**Number of persons aged 65 or older in the household**
HHCOMP	**Households composition;** 10 – *One adult without children* *One adult with at least*: 11 – an child aged less than 15 12 – else: an child aged 15 to 24 20 – *One couple without children* *One couple with at least*: 21 – an child aged less than 15 22 – else: an child aged 15 to 24 30 – Two adults (not a couple) or more without children *Two adults (not a couple) or more with at least:* 31 – an child aged less than 15 32 – else: an child aged 15 to 24 50 – Blank

Code	Definition
HHNBWORK	**Number of employed persons in the household**
HHNBEMPL	**Number of employed adults in the household**
HHNBUNEM	**Number of unemployed adults in the household**
HHNBINAC	**Number of inactive adults in the household**
HHPERSnew	**Persons aged 15–24 living with their family** 1 – Inactive 15–24 aged living with family 0 – otherwise
HHNBCHLDnew	**Number of children in the household** 0 – No children in the household 1 – One children in the household 2 – Two children in the household 3 – Three children in the household 4 – Four or more children in the household
HATLEVnew	**Highest educational attainment level** 0 – No formal education 1–6 ISCED level of education
HATFIELDnew	**Field of education** 0 – General programs 1 – Education 2 – Humanities, languages and arts 3 – Social sciences, business and law 4 – Natural sciences, mathematics and statistics 5 – Information and communication technologies 6 – Engineering, manufacturing and construction 7 – Agriculture, forestry, fisheries and veterinary 8 – Health and welfare 9 – Services 10 – Unknown or unspecified 999 – Not applicable
EDUCLEVLnew	**Level of education of student or apprentice in regular education during the last 4 weeks** 1–6 ISCED level of education 9 – Not applicable (Has not been a student or apprentice)
REGIONnew	**Region of household (NUTS2 classification)** 31 – Northwest 32 – North Central 33 – Northeast 34 – Southeast 41 – Southwest 42 – South Central

B Appendix

B.1 Labour Force Participation: Descriptive Statistics and Evolution across Time and Subsets

Figure 2: **Number of respondents with active and non-active labour market status**

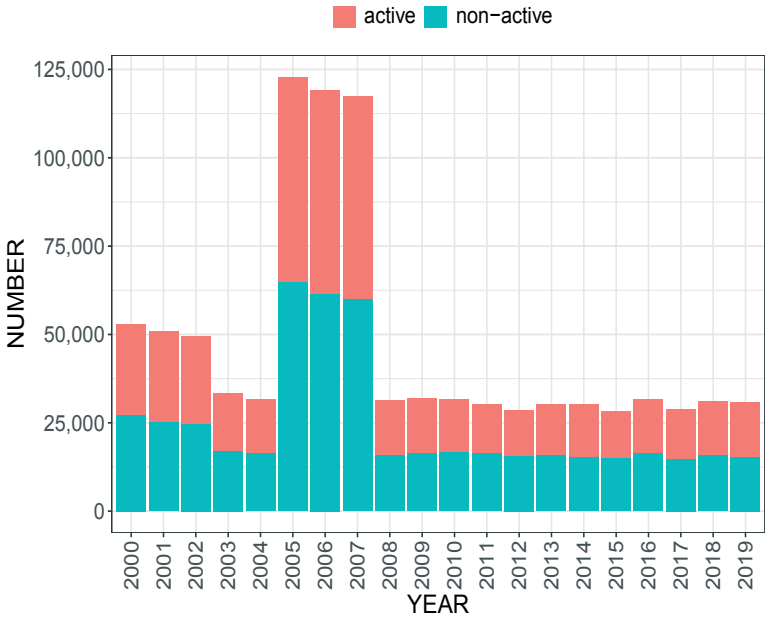
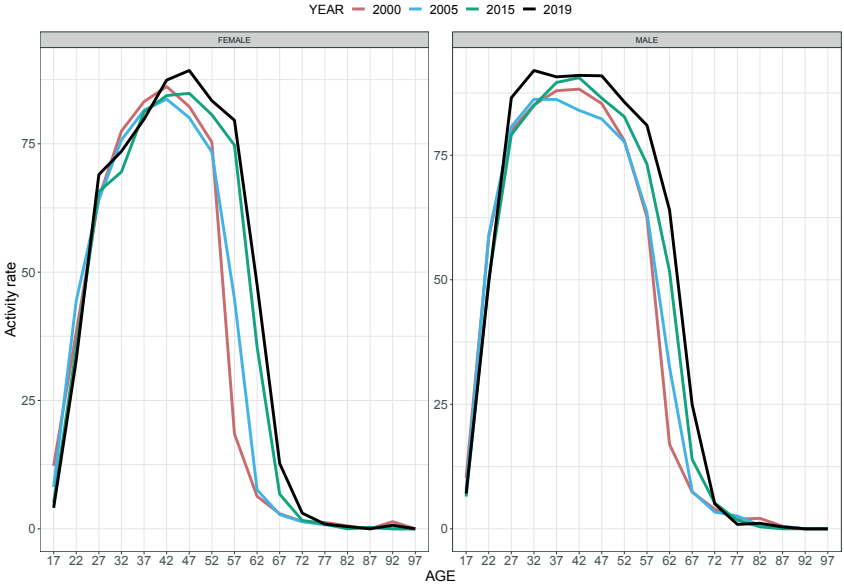


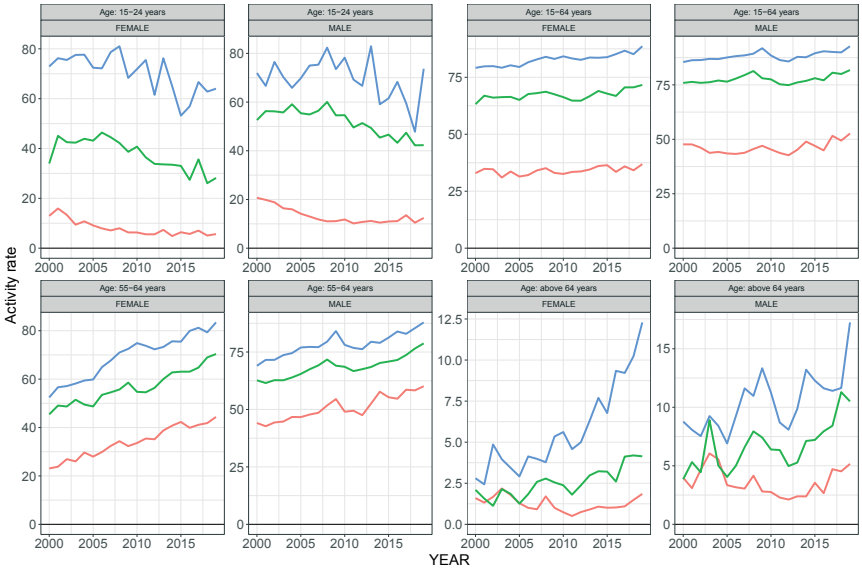
Figure 3: **Micro-based activity rates *versus* LFS macro data**

Figure 4: **Activity rates, unweighted: age groups**



Note: The figure presents activity rates by the following age groups:
17 (age group 15–19 years),
22 (age group 20–24 years),
27 (age group 25–29 years),
32 (age group 30–34 years),
37 (age group 35–39 years),
42 (age group 40–44 years),
47 (age group 45–49 years), *etc.*

Figure 5: **Activity rates, unweighted: education level I**



Note: Education level is defined as H:high, M:medium, L:low.

Figure 6: **Activity rates, unweighted: degree of urbanisation**

Note: Degree of urbanisation:
1 Cities (Densely populated area),
2 Towns and suburbs (Intermediate density area),
3 Rural area (Thinly populated area).

Figure 7: **Activity rates in the 20–49 age group, unweighted: marital status**

Figure 8: **Activity rates in the 20–49 age group, unweighted: number of children in the family**

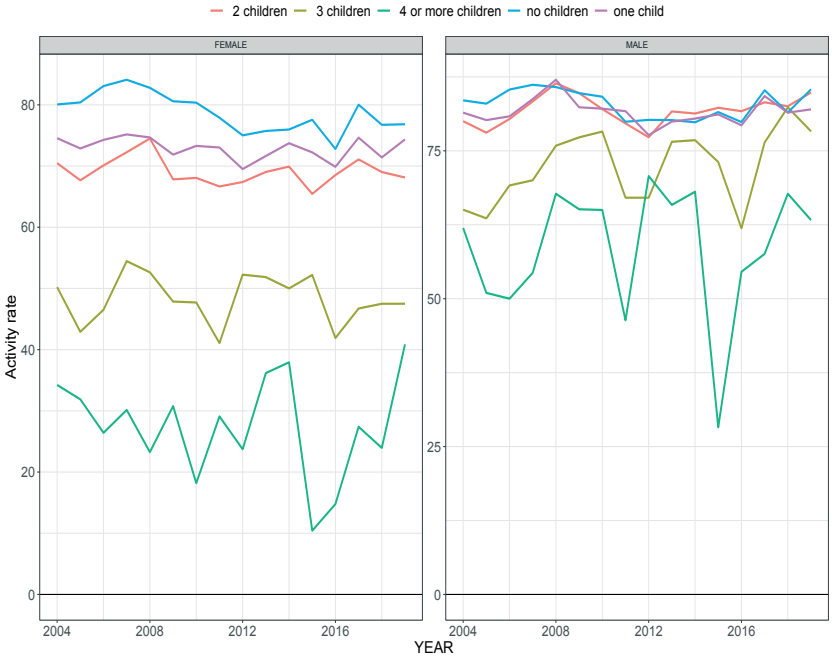


Figure 9: **Activity rates in the 20–49 age group, unweighted: age of the children in the family**

Figure 10: **Activity rates in the 20–49 age group, unweighted: number of elderly people living in the household**

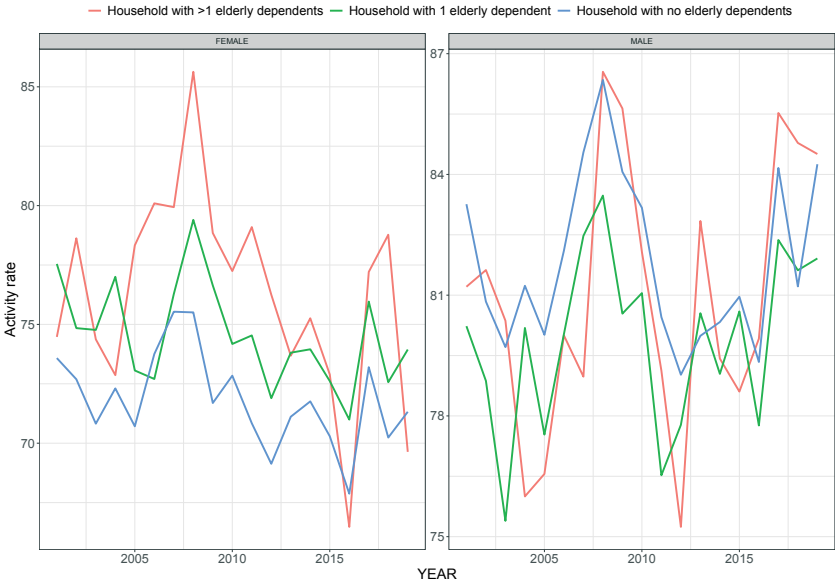
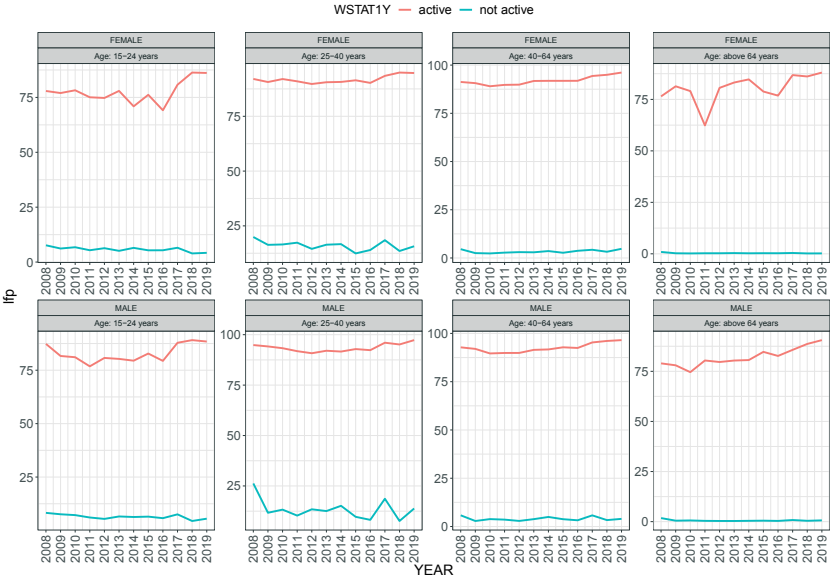


Figure 11: **Activity rates, unweighted: activity one year ago**

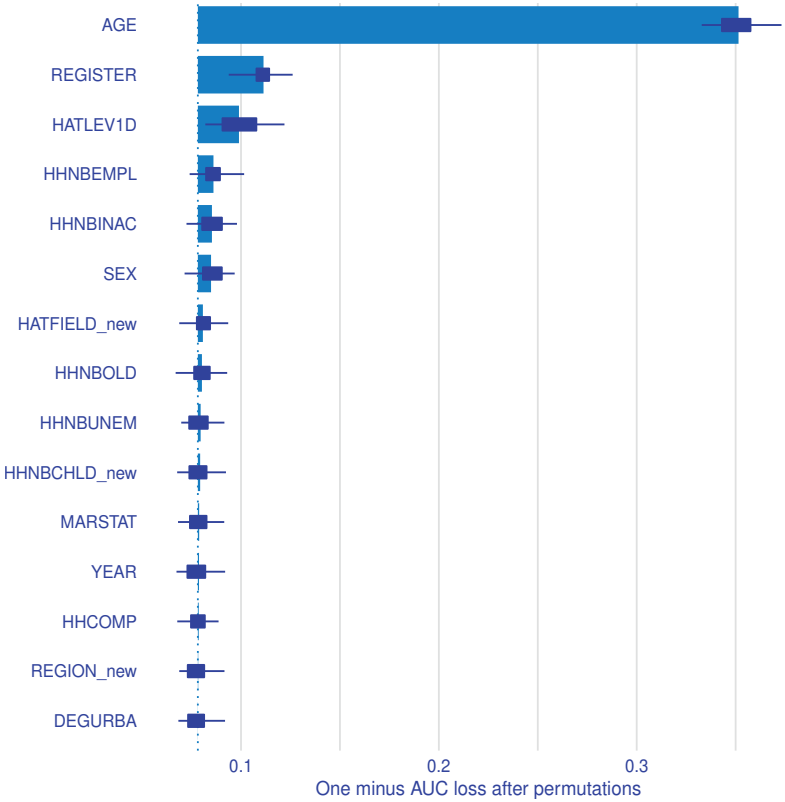


C Appendix

C.1 Selected Results from the ML Models

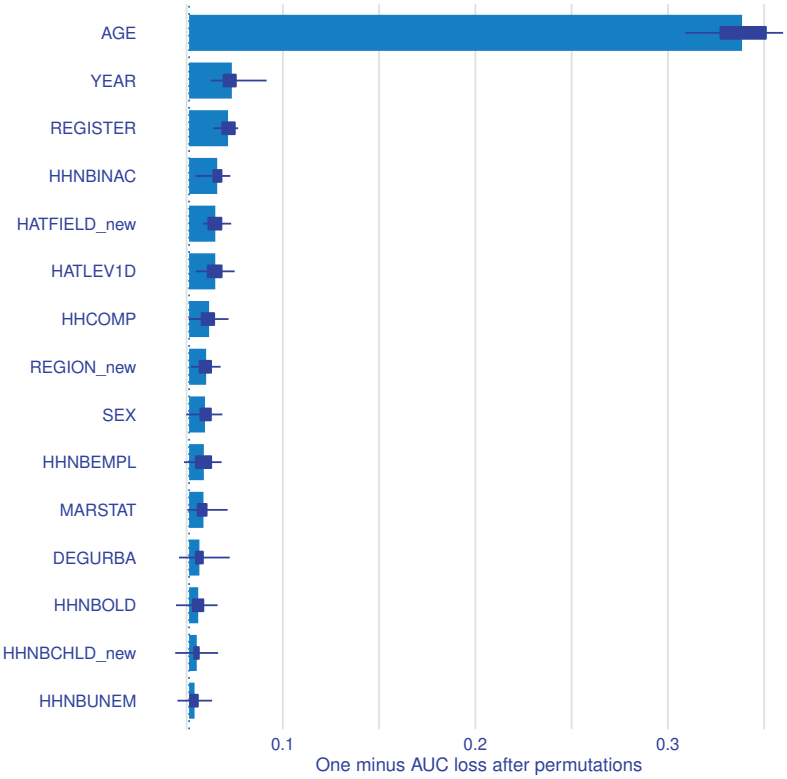
Note: The precise definitions of the variables appearing in the figures are presented in Appendix A, Table 4.

Figure 12: **Permutation importance plot for the elastic net model.**



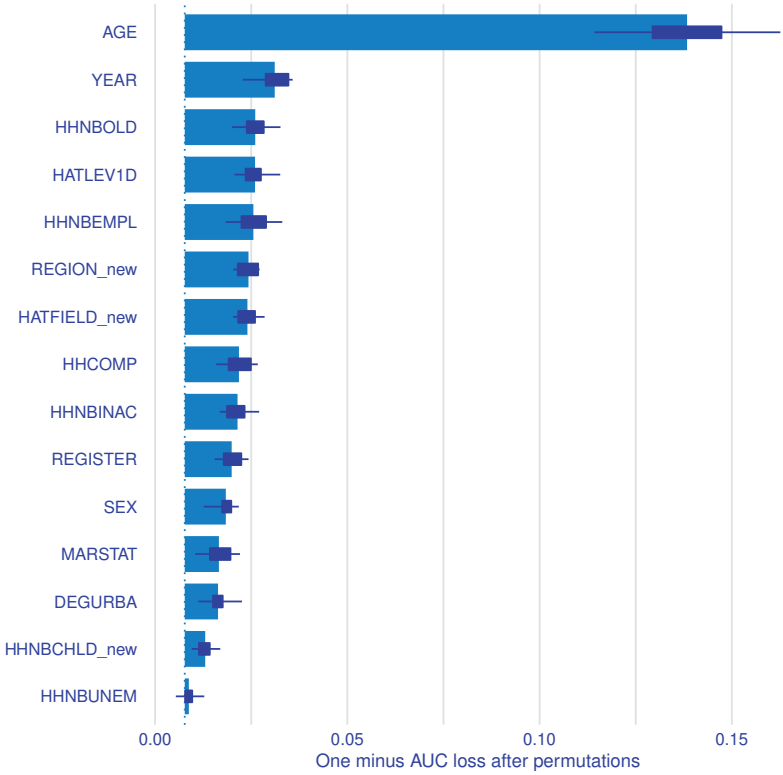
Variables: AGE – Age of interviewed person, REGISTER – Registration at a PE office, HATLEV1D – Level of education, HHNBEMPL – Number of employed adults, HHNBINAC – Number of inactive adults, SEX – Sex of the respondent, HATFIELD_new – Field of education, HHNBOLD – Number of persons aged 65, HHNBUNEM – Number of unemployed adults, HHNBCHLD_new – Number of children, MARSTAT – Marital status, YEAR – Fixed reference year, HHCOMP – Households composition, REGION_new – Region of household, DEGURBA – Degree of urbanisation.

Figure 13: **Permutation importance plot for the SVM model**



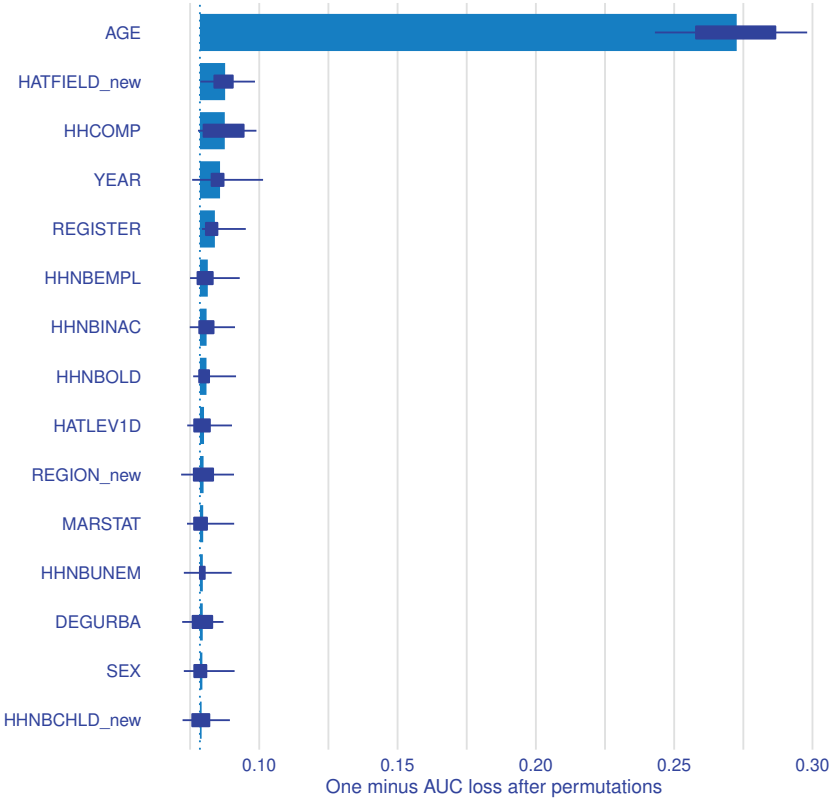
Variables: AGE – Age of interviewed person, YEAR – Fixed reference year, REGISTER – Registration at a PE office, HHNBINAC – Number of inactive adults, HATFIELD_new – Field of education, HATLEV1D – Level of education, HHCOMP – Households composition, REGION_new – Region of household, SEX – Sex of the respondent, HHNBEMPL – Number of employed adults, MARSTAT – Marital status, DEGURBA – Degree of urbanisation, HHNBOLD – Number of persons aged 65, HHNBCHLD_new – Number of children, HHNBUNEM – Number of unemployed adults.

Figure 14: **Permutation importance plot for the random forest model**



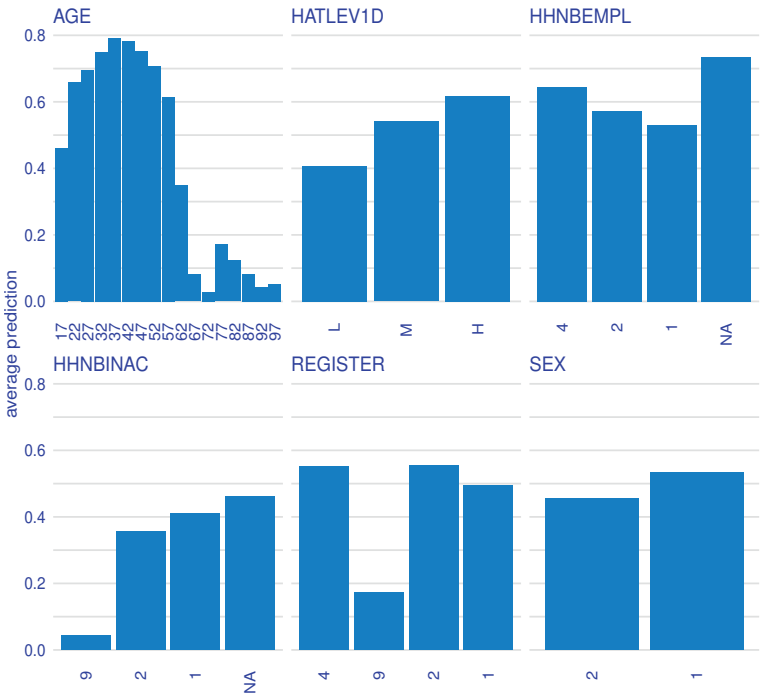
Variables: AGE – Age of interviewed person, YEAR – Fixed reference year, HHNBOLD – Number of persons aged 65, HATLEV1D – Level of education, HHNBEMPL – Number of employed adults, REGION_new – Region of household, HATFIELD_new – Field of education, HHCOMP – Households composition, HHNBINAC – Number of inactive adults, REGISTER – Registration at a PE office, SEX – Sex of the respondent, MARSTAT – Marital status, DEGURBA – Degree of urbanisation, HHNBCHLD_new – Number of children, HHNBUNEM – Number of unemployed adults.

Figure 15: **Permutation importance plot for the KNN model**



Variables: AGE – Age of interviewed person, YEAR – Fixed reference year, HHNBOLD – Number of persons aged 65, HATLEV1D – Level of education, HHNBEMPL – Number of employed adults, REGION_new – Region of household, HATFIELD_new – Field of education, HHCOMP – Households composition, HHNBINAC – Number of inactive adults, REGISTER – Registration at a PE office, SEX – Sex of the respondent, MARSTAT – Marital status, DEGURBA – Degree of urbanisation, HHNBCHLD_new – Number of children, HHNBUNEM – Number of unemployed adults.

Figure 16: **Partial dependence plots for selected variables from the elastic net model**



Variables: AGE – Age of interviewed person, HATLEV1D – Level of education, HHNBEMPL – Number of employed adults, HHNBINAC – Number of inactive adults, REGISTER – Registration at a PE office, SEX – Sex of the respondent.

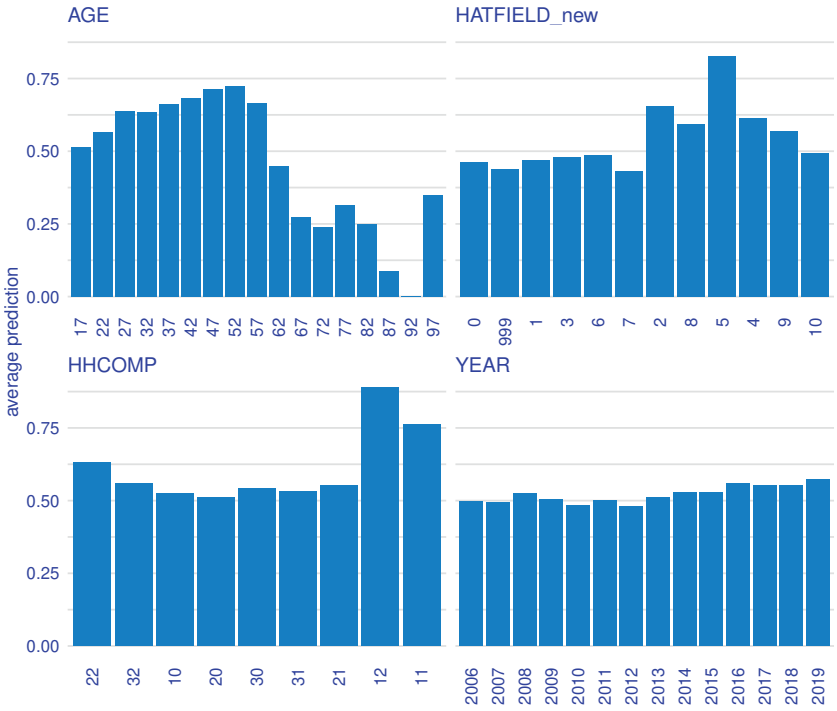
Figure 17: **Partial dependence plots for selected variables from the SVM model**

Variables: AGE – Age of interviewed person, HATFIELD_new – Field of education, HATLEV1D – Level of education, HHNBINAC – Number of inactive adults, REGISTER – Registration at a PE office, YEAR – Fixed reference year.

Figure 18: **Partial dependence plots for selected variables from the random forest model**

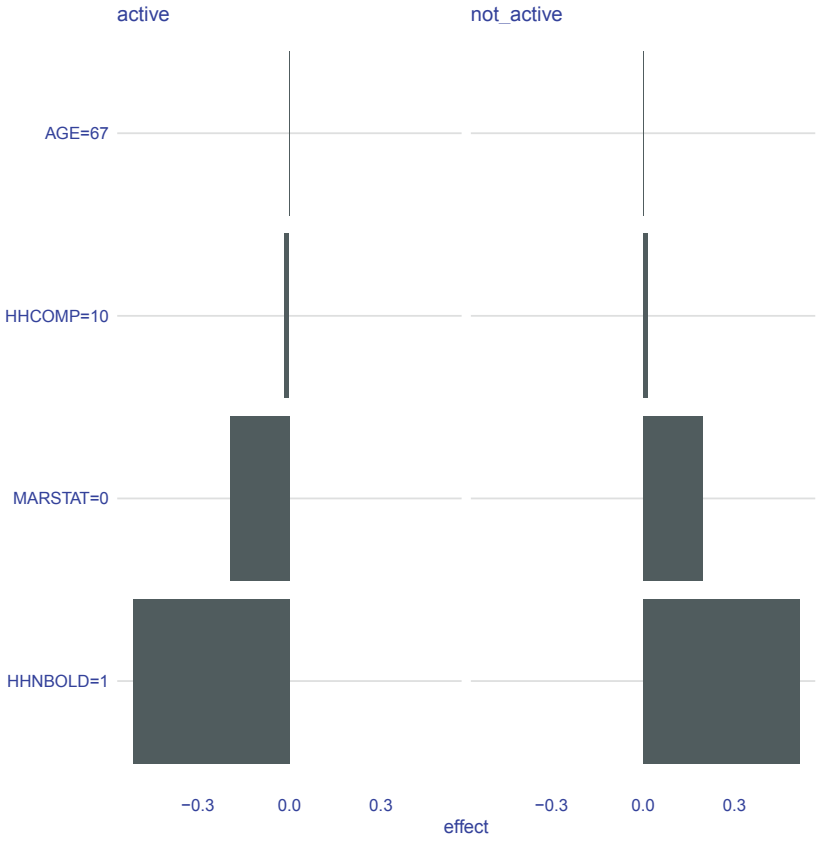
Variables: AGE – Age of interviewed person, HATFIELD_new – Field of education, HATLEV1D – Level of education, HHNBEMPL – Number of employed adults, HHNBOLD – Number of persons aged 65, REGION_new – Region of household, YEAR – Fixed reference year.

Figure 19: **Partial dependence plots for selected variables from the KNN model**



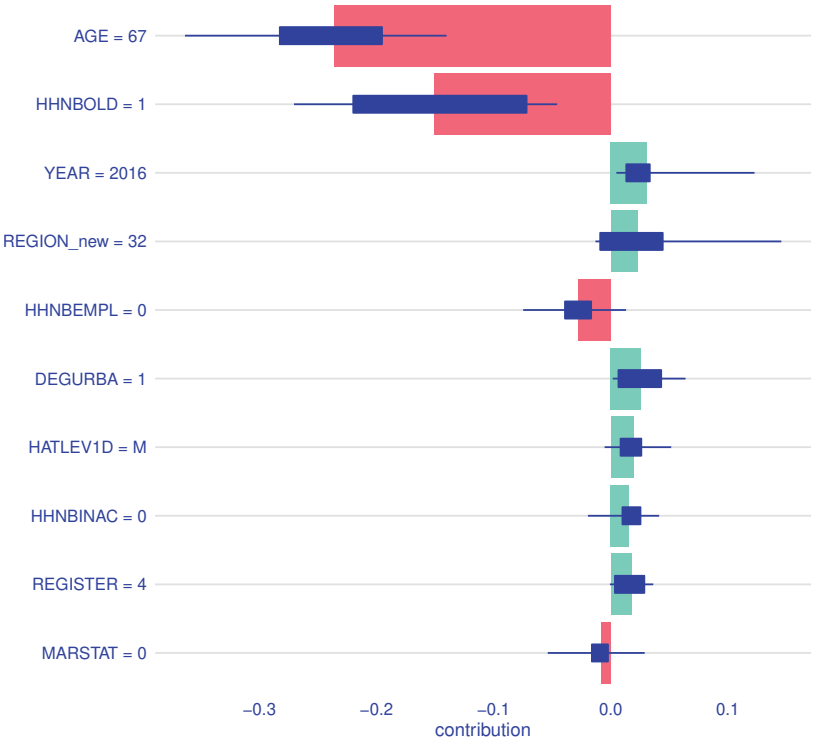
Variables: AGE – Age of interviewed person, HATFIELD_new – Field of education, HHCOMP – Households composition, YEAR – Fixed reference year.

Figure 20: **A LIME explanation of a prediction generated by the KNN model**



Variables: AGE – Age of interviewed person, HHCOMP – Households composition, MARSTAT – Marital status, HHNBOLD – Number of persons aged 65.

Figure 21: **A Shapley value explanation of a prediction generated by the random forest model**



Variables: AGE – Age of interviewed person, HHNBOLD – Number of persons aged 65, YEAR – Fixed reference year, REGION_new – Region of household, HHNBEMPL – Number of employed adults, DEGURBA – Degree of urbanisation, HATLEV1D – Level of education, HHNBINAC – Number of inactive adults, REGISTER – Registration at a PE office, MARSTAT – Marital status.

ISBN 978-619-7409-26-0 (online)

The sculptural composition by Kiril Shivarov depicting Hermes and Demeter on the southern façade of the Bulgarian National Bank building is used in cover design