Term Paper Proposal

Sarah Gust *ifo Institute at the University of Munich, gust@ifo.de*
Ann-Christin Kreyer *Max Planck Institute for Innovation and Competition, ann-christin.kreyer@ip.mpg.de*

This is where we put the abstract...

Introduction

Digitization has fundamentally changed labor demand. Middle-skilled workers were replaced by technology while the demand for high skilled workers and low-skilled workers has grown (Acemoglu and Autor, 2011; Autor and Dorn, 2013). Autor and Dorn (2013) show that increased employment at the lower tail of the earnings distribution is mainly due to an increase in service occupations. At the upper tail technological changed led to a college wage premium: wages of college graduates relative to high-school graduates increased (Acemoglu and Autor, 2011). De La Rica et al. (2020) develop abstract, routine and manual task measures and find that a one-standard-deviation increase in abstract tasks is related to a 3.3-log-point wage premium. For each standard deviation of routine tasks there is a 2.6 to 2.9-log-point wage penalty. The vast literature on the change of job tasks takes skills of workers as pre-defined. However, workers and firms can also invest in new skills via training. Especially the emergence of Massiv Open Online Courses (MOOC) over the past years has facilitated global access to ICT and programming courses. In this paper, we investigate the specific characteristics of workers that participate in on-the-job training and open education. Is training and especially new open educational programs an opprtunity for middle skilled workers that are primarily effected by decreasing job opportunities to take on more abstract tasks? Or does training reinforce inequalities because only high-skilled workers receive and invest in training?

Becker (1962) distinguished between two kinds of training: specific and general. Specific training increases the marginal product of a worker within one specific firm while general training increases her productivity in many other firms. In a perfect labor market workers are paid their marginal product. In such case, firms would not invest into general training of their employees as they could leave the firm and look for a better paid job. Instead, workers would pay for their general training as an investment into higher future wages. Lynch (1991) and Lynch (1992) find that on-the job training tends to be firm specific in the US and thus wage raises cannot be taken along to subsequent employers. Off-the-job training by proprietary institution have little effect on wages in the current employment but raise future expected wages in subsequent employment. Acemoglu and Pischke (1999) argue that firms still invest in general training due to their monopsony power. Wages increase by less than the marginal productivity and firms can profit. Konings and Vanormelingen (2015) find that an increase in the share of trained workers by 10 percentage points raises the productivity by 1.7 to 3.2 percent while wages only increase by 1.0 to 1.7 percent.

Previous literature on training focuses on wage and productivity effects but the research on the specific characteristics of workers that participate in training is scarce. Applying a machine learning, we can identify the factors that drive the probability of receiving training from a large set of ## variables of the survey of the Programme for the International Assessment of Adult Competencies (PIAAC). Morespecifically (what we do)

- What do we find
- Further research
- mostly effect on wages, inequality

Data and Desriptive Statistics

To explore these questions we use the results of the survey of the Programme for the International Assessment of Adult Competencies (PIAAC) (GESIS - Leibniz Institute for the Social Sciences (n.d.)). The survey was conducted by the Organisation for Economic Co-operation and Development (OECD) with the goal to assess which skills adults need to manage challenges and tasks at work as well as in their personal life. The study targeted explicitly the skills in literacy, numeracy and adaptive problem solving while also comprising comprehensive background information on the respondents past and current education, subjective assessments of their skills and job requirements as well as information on migration (GESIS - Leibniz Institute for the Social Sciences (n.d.)). The first cycle consisted of three rounds and began in 2011/12. In the first round 24 countries took part. In the second round nine additional countries participated and in the last round individuals from five different countries were questioned. In total 40 countries participated in the first cycle comprising about 5,000 randomly selected adults who were between 16 and 65 years old. The second cycle started in 2018 and results are to be expected in 2022 (GESIS - Leibniz Institute for the Social Sciences (n.d.)).

For the study at hand the results of the first wave are used in a reduced form. The original dataset comprises 1,460 columns with 230,691 observations of respondents. However, the 'research question of this paper is to analyze the probability of trainings for middle-skilled workers compared to trainings for high-skilled and low-skilled workers in the wake of the increasing polarization of skills following from digitization'. To answer this, the original PIAAC dataset is reduced to 130 variables in total, including indices. For example, information on the various test results conducted in the study are excluded. To achieve comparability across countries, questions which were only answered by respondents living in the United States are also excluded. The final dataset is cross-sectional with one observation representing the answers of one respondent.

The 130 variables we kept in the final dataset comprise information on the individual's background information, her past and ongoing formal or informal education, information on training activities, information on ICT skills and the respective extensive and intensive margin, her subjective job requirements, information on her curent job and information on monthly income. As we do not restrict the dataset in terms of respondents but only in terms of questions answered, our final dataset comprises 230,691 observations of individuals. Of those 230,691 individuals, 122,830 are female and 107,859 are male (see Figure 1). The age of the respondents is evenly distributed between the ages 16 to 64 with a female mean age of 39.95 years and male mean age of 39.38 (see Figure 2).

The classification of the respondents jobs in terms of skills is also evenly distributed across age groups and gender. However, respondents working in semi-skilled white-collar occupations are slightly younger than those working in skilled occupations or semi-skilled blue-collar occupations (see Fiugre 3).

The key variables of this study are the the skill classification of the individual's job and her trainings comprising on-the-job training, seminars or workshops, distance or open training courses as well as private lessons. The simple OLS regressions reveal that there are positive correlations between the high-skilled jobs and the number of trainings respondents participated in. This holds true for on-the-job-training, seminars or workshops, distance or open educational training as well as for private lessons. However, for semi-skilled jobs, the picture is slightly different. Here, the number of seminars or workshops and private lessons are positively correlated with the semi-skilled occupations.

Figure 1: Distribution of Gender

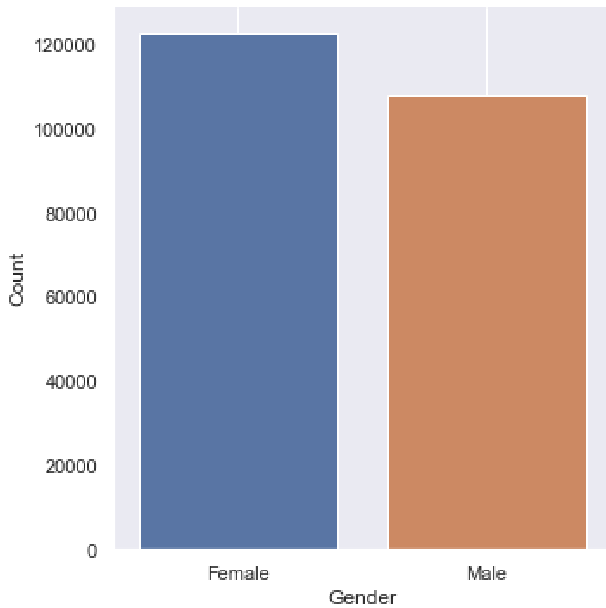
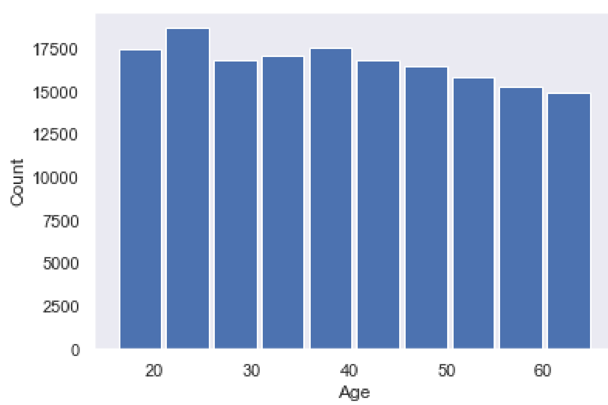
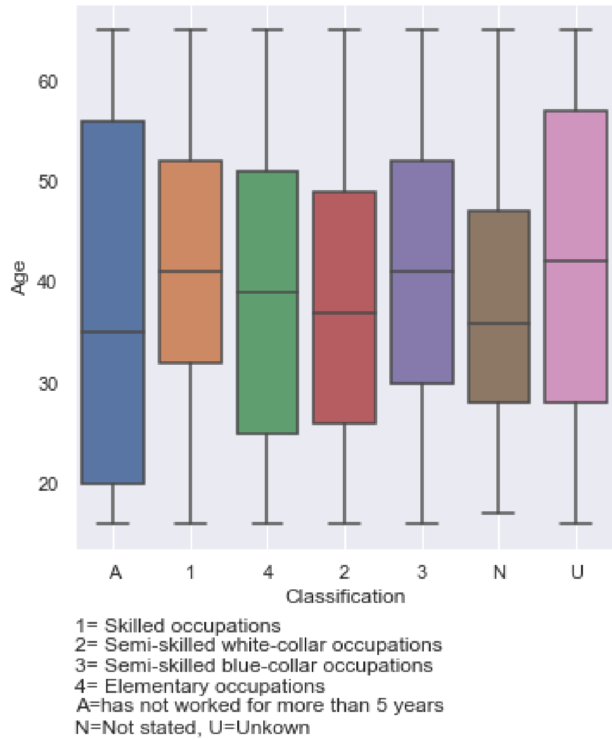


Figure 2: Distribution of Age



- Name, source, unit, time, structure, number of observations, relevant population.
- Definition of (main) sample.
- Definition and characteristics of key variables.
- Limitations and potential biases.
- Provide the data and the software code (replication).

Figure 3: Age and Classification of jobs



- Plot the main empirical associations you want to study!
- Do NOT assume the reader knows anything about these data

Lasso Linear Model

We start with a simple linear regression model $y = X\beta + \varepsilon$, where $y \in \mathbb{R}^N$ is the predicted participation in on-job or off-job training, $X \in \mathbb{R}^{N\times k}$ are the vectors of covariates, and $\varepsilon \in \mathbb{R}^N$ is the residual with the standard assumptions of OLS. To select the set of covariates with the strongest predictive power from our set of $\ldots$ variables, we apply the Least absolute shrinkage and selection operator (Lasso) that was first proposed by Tibshirani (1996). We add the Lasso penalty equal to $\sum_k |\beta_k|$ to our linear model. The Lasso linear estimator $\hat{\beta}$ is then given by

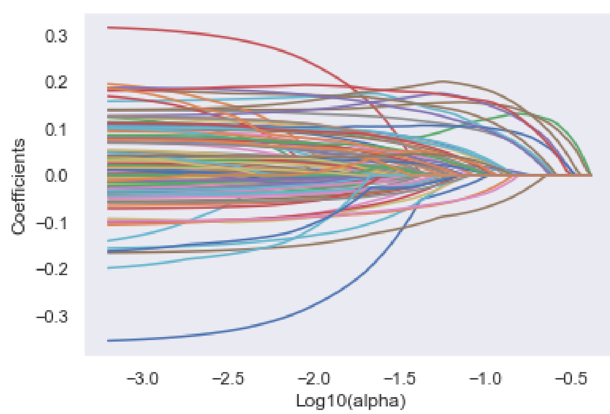
$$\hat{\beta}_\lambda = argmin\{\sum_i (y_i - x_i^T\beta)^2 + \lambda \sum_k |\beta_k|\} \quad (1)$$

That is, we minimize the sum of squared residuals but the Lasso penalty puts a cost at every $\hat{\beta} \neq 0$ and thus, we penalize complexity and avoid over-fitting the model. $\lambda > 0$ is the penalty weight or the *tuning parameter*.

Training the Model

Figure 4 and Figure 5 show the lasso regularization path of candidate models $\hat{\beta}_1 \ldots \hat{\beta}_t$ that we obtained by minimizing Equation 1 over a sequence of tuning parameters $\lambda_1 < \lambda_2 < ... < \lambda_T$ with on-job training and off-job training as dependen variable respectively. The vertical axis contains different levels of $\hat{\beta}$. The horizontal axis contains different levels for λ. Each vertical section along the horizontal axis, represents one candidate model. Moving from higher to lower λ, the algorithm includes more nonzero $\hat{\beta}_k$ and becomes more complex. To find the optimal value for λ we use *5-fold cross validation*. We split the data in 5 random evenly sized subset and derive the lasso paths $\hat{\beta}_1^k \ldots \hat{\beta}_T^k$ on each of the folds but one fold to train the models. Then we use the left out fold to obtain the out-of-sample error for each candidate model. The best $\hat{\lambda}_t$ minimizes the out-of-sample error.

Figure 4: Lasso path for on-job training



describe optmal λ and p* here

Variable Selection

present most important variables here

Out-of-Sample Evaluation Results

present out of sample performance and compare to OLS

Figure 5: Lasso path for off-job training

Lasso Logistic Model

We now estimate the Lasso model for the binary oucome variables of whether or not a person received on-job training and whether or not a person participated in an off-job training. We estimate the Lasso-regularized logistic model for the probability that a person received one specific training as follows:

$$\hat{\theta}_\lambda = argmin(-l_N(\theta)) + \lambda \sum_k |\theta^k| \tag{2}$$

where $l_N(\theta))$ is the log-likelihood function

$$l_N(\theta)) = \sum_i [y_i x_i \theta - log(1 + e^{x_i \theta})]$$

$\sum_k |\theta^k|$ is the lasso penalty that shrinks coefficients of little explanatory power to zero. $\lambda > 0$ is the penalty weight.

Training the Model

Figure 6 and Figure 7 present the Lasso regularization path for the logistic candidate models. The models are ordered from the most penalized to the least penalized model and the algorithm includes more non-zero coefficients in the model. We select λ via *5-fold cross validation* which leads to an optimal λ of 29.764 for on-job training and an optimal λ of 0.089 for open education.[1]

[1] Note that these results include randomization which may lead to different outcomes if run again.

Figure 6: Logistic model: Lasso path for on-job training

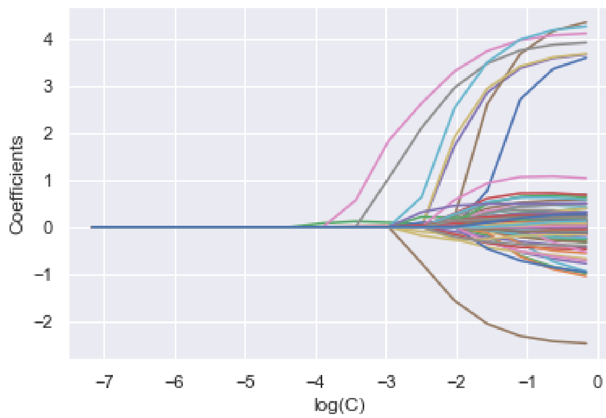
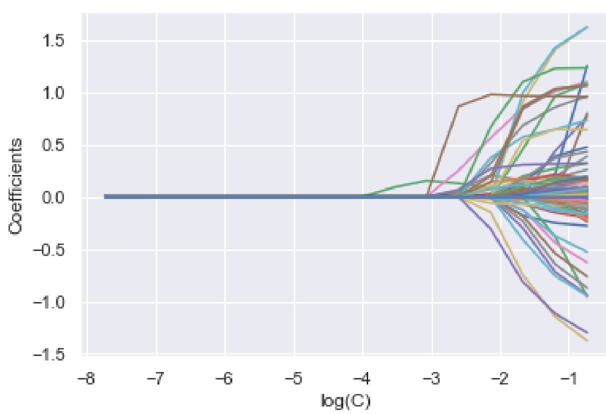


Figure 7: Locistic model: Lasso path for off-job training



Variable Selection

The Lasso logit model indentifies 152 columns[2] with non-zero predictive power for on-job training. We present the coefficients in Table . Let us first turn to the results for the skill level. *skill_4* are elementary occupations and represent the reference group here. *skill_3* is the dummy variable for semi-skilled blue-collar occupations, *skill_2* is the dummy for semi-skilled white-collar occupations, and *skill_1* is the dummy for skilled occupations. Working in a semi-skilled blue-collar occupation and working in a semi-skilled white-collar occupation increases the probability of receiving training by 00.45 % and 00.42 % respectively compared to working in an elemetary occupation. Working in a skilled occupation increases the probability of receiving training by 10.11 % relative to working in an elementary occupation.

The dummy variable for whether a person was employed during studying for a qualification, *b_q10a_Yes*, has the highest positive explanatory power. If a person uses a computer on this specific job (*g_q04_Yes*), it increased the probability of receiving on-job training by 67.52 %. Having general computer experience (*computerexperience_Yes*) increases the probability of receiving on-job training by 38.93 %. People are 24.52 % more likely to participate in on-job training, if they have the feeling that they need more training in order to cope well with their present duties (*f_q07b_Yes*). Moreover, employees are more likely to receive training if they work in larger companies, compared to smaller companies and if they have a higher educational level.

If a job does not involve keeping up to date with new services and products (*d_q13c_Never*), it lowers the probability of reveiving training by 37.60 %. Never participating in online discussions such as conferences (*g_q05h_Never*) reduces the probability of receiving on-job training by 39.00 %. Also, having a low education level, if a job needs less than one month of prior work experience, and working in a job without a contract have the most negative effects on the chances of participating in on-job training.

For off-job training, the Lasso logit model identifies 138 non-zero columns[3] The results look very similar to the on-job training. Working in a skilled occupation increases the probability of participating in open education by 11.03 %. The indicator for semi-skilled blue-collar or white-collar workers zero and thus excluded by the Lasso regularization. As in on-job training, the

[2]Including country and industry controls
[3]Including country and industry controls

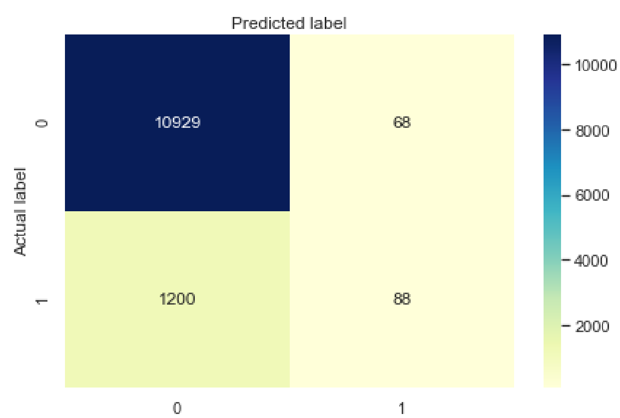
most important feature is being employed and using the computer on this specific job (66.60 %) or having general computer experience (33.45 %). Also working in a larger company and having a higher educational level increases the probability of participating in open education. Lower education, having no contract, and never using the computer for work tasks negatively effect the probability of participating in open education.

- explain implications and relate to literature

Accuracy of the Model

Finally, we evaluate the accuracy of the Lasso logit model. Figure 8 presents the confusion matrix for on the job training. 10929 are true negative prediction and there are 88 true positive predictions. The model yields 1200 false negative predictions and 68 false positive predictions. We present the confision matrix for open education in Figure 9. For this model, we obtain 10965 true positive predictions and 64 true negative predictions. 1224 + 32 are incorrect predictions.

Figure 8: Logistic model: Confusion matrix for on-job training



In Table 1, we compare the accuracy of the Lasso logistic model with the unregularized logistic model. The test accuracy of the logistic model is 0.4858 and the test accuracy of the Lasso logistic model is 0.7292 for on-job training. The test accuracy for open education is 0.5021 for the logistic model and reaches 0.8188 with the Lasso logistic model. With the Lasso penalty we excluded unnecessary variables from our Logistic regression that caused over-fitting. The Lasso-regularized logistic model performs much better for both outcome variables.

Figure 9: Locistic model: Confusion matrix for off-job training

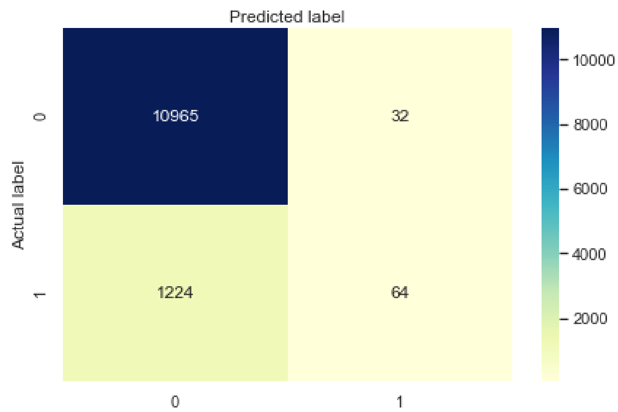


Table 1: Accuracy of the Lasso logistic model

	Lasso Logistic Model	Logistic Model
On-job training		
Training accuracy	0.7288	0.4962
Test accuracy	0.7292	0.4858
Open education		
Training accuracy	0.8185	0.4996
Test accuracy	0.8188	0.5021

- Describe how the hypothesis is linked to your estimation.
- Describe the estimation using equations.
- Discuss the parameters and variables.
- What are the identifying assumptions, what are (possible) violations and their consequences?
- What will you do about this?
- Tell a story!
- Guide the reader.
- Focus on the key points, not the details.
- Discuss quality and quantity.
- Discuss problems.

(APPENDIX) Appendix

Appendix: Tables

Table: Lasso logistic regression for on-job training

	Coefficients	Feature
0	-0.0173091	age_r
1	0.00648863	j_q03b
2	0.0323998	yrsget
3	0.014883	c_q09
4	0.00407624	c_q10a
5	0.0688914	readytolearn
6	-4.95059e-08	earnmthallppp
7	0.389253	computerexperience_Yes
8	0.152316	d_q09_A temporary employment agency contract
9	-0.0685864	d_q09_An apprenticeship or other training scheme
10	0.000118794	d_q09_An indefinite contract
11	-0.249488	d_q09_No contract
12	0.268225	d_q09_Other
13	0.0919625	f_q07a_Yes
14	-0.00312127	b_q01b_Engineering, manufacturing and construction
15	-0.079488	b_q01b_General programmes
16	0.192744	b_q01b_Health and welfare
17	-0.0830876	b_q01b_Humanities, languages and arts
18	0.0623138	b_q01b_Science, mathematics and computing
19	-0.0454431	b_q01b_Services
20	-0.00709438	b_q01b_Social sciences, business and law
21	0.0955665	b_q01b_Teacher training and education science
22	0.0628176	d_q06b_Increased
23	-0.0100635	d_q06b_Stayed more or less the same
24	0.0822428	d_q04_t_Employee, supervising fewer than 5 people
25	0.112577	d_q04_t_Employee, supervising more than 5 people
26	-0.162508	g_q08_Yes
27	-0.119125	pared_At least one parent has attained tertiary
28	-0.0624561	pared_Neither parent has attained upper secondary

	Coefficients	Feature
29	-0.0492441	gender_r_Male
30	-0.655047	leaver1624_Not in education, did not complete ISCED 3, aged 16 to 24
31	-0.00587444	d_q13c_Every day
32	-0.132742	d_q13c_Less than once a month
33	-0.159028	d_q13c_Less than once a week but at least once a month
34	-0.375997	d_q13c_Never
35	-0.160666	j_q04a_Yes
36	0.343026	edcat8_Post-secondary, non-tertiary (ISCED 4A-B-C)
37	-0.162479	edcat8_Primary or less (ISCED 1 or less)
38	0.128055	edcat8_Tertiary - bachelor/master/research degree (ISCED 5A/6)
39	0.350122	edcat8_Tertiary – bachelor degree (ISCED 5A)
40	0.31369	edcat8_Tertiary – master degree (ISCED 5A)
41	0.165938	edcat8_Tertiary – professional degree (ISCED 5B)
42	-0.152984	edcat8_Tertiary – research degree (ISCED 6)
43	0.0797353	edcat8_Upper secondary (ISCED 3A-B, C long)
44	0.160175	g_q07_Yes
45	0.00541096	vet_True
46	0.100965	g_q05d_Every day
47	-0.104847	g_q05d_Less than once a month
48	0.0893642	g_q05d_Less than once a week but at least once a month
49	0.0191254	g_q05d_Never
50	-0.308108	d_q14_Extremely dissatisfied
51	-0.00958417	d_q14_Extremely satisfied
52	-0.104017	d_q14_Neither satisfied nor dissatisfied
53	-0.0803196	d_q14_Satisfied
54	0.0423771	g_q05a_Every day
55	-0.0825071	g_q05a_Less than once a month
56	0.108813	g_q05a_Less than once a week but at least once a month
57	-0.0223103	g_q05a_Never
58	0.153748	g_q05e_Every day
59	0.10127	g_q05e_Less than once a month
60	0.118944	g_q05e_Less than once a week but at least once a month
61	0.162112	g_q05e_Never
62	0.0906379	g_q05f_Every day
63	-0.0120091	g_q05f_Less than once a month

	Coefficients	Feature
64	-0.105233	g_q05f_Less than once a week but at least once a month
65	-0.0498476	g_q05f_Never
66	-0.151607	g_q05g_Every day
67	0.0534463	g_q05g_Less than once a month
68	0.0319572	g_q05g_Less than once a week but at least once a month
69	-0.0987864	g_q05g_Never
70	0.00587424	g_q05h_Every day
71	-0.00152421	g_q05h_Less than once a month
72	-0.0529479	g_q05h_Less than once a week but at least once a month
73	-0.390016	g_q05h_Never
74	0.245238	f_q07b_Yes
75	-0.125719	b_q10c_Not useful at all
76	0.431681	b_q10c_Somewhat useful
77	-0.0221019	b_q10c_Very useful
78	-0.0368781	d_q12c_1 to 6 months
79	-0.0163393	d_q12c_3 years or more
80	-0.137065	d_q12c_7 to 11 months
81	-0.277454	d_q12c_Less than 1 month
82	0.0594719	d_q12c_None
83	0.0659232	g_q05c_Every day
84	-0.126711	g_q05c_Less than once a month
85	-0.0997177	g_q05c_Less than once a week but at least once a month
86	-0.250353	g_q05c_Never
87	-0.123279	d_q12b_A lower level would be sufficient
88	-0.177648	d_q12b_This level is necessary
89	0.111568	d_q06a_11 to 50 people
90	0.280167	d_q06a_251 to 1000 people
91	0.22403	d_q06a_51 to 250 people
92	0.293531	d_q06a_More than 1000 people
93	-0.277048	d_q03_The private sector (for example a company)
94	-0.1113	d_q03_The public sector (for example the local government or a state school)
95	0.675257	g_q04_Yes
96	0.801725	b_q10a_Yes
97	0.0374759	g_q06_Moderate
98	-0.0442568	g_q06_Straightforward

	Coefficients	Feature
99	0.101065	skill_1
100	0.00415191	skill_2
101	0.00447484	skill_3

Table: Lasso logistic regression for off-job training

	Coefficients	Feature
0	-0.0176953	age_r
1	0.0474509	yrsget
2	0.0131248	c_q09
3	0.00448319	c_q10a
4	0.0861567	readytolearn
5	-3.07033e-08	earnmthallppp
6	0.220386	computerexperience_Yes
7	-0.21247	d_q09_No contract
8	0.0751546	f_q07a_Yes
9	-0.0165669	b_q01b_General programmes
10	0.179973	b_q01b_Health and welfare
11	0.0139165	b_q01b_Science, mathematics and computing
12	0.0940462	b_q01b_Teacher training and education science
13	0.0570971	d_q06b_Increased
14	-0.00782241	d_q06b_Stayed more or less the same
15	0.0140561	d_q04_t_Employee, supervising fewer than 5 people
16	0.0829086	d_q04_t_Employee, supervising more than 5 people
17	-0.0656784	g_q08_Yes
18	-0.042624	pared_At least one parent has attained tertiary
19	-0.024773	gender_r_Male
20	0.048559	d_q13c_Every day
21	-0.0795516	d_q13c_Less than once a month
22	-0.0931521	d_q13c_Less than once a week but at least once a month
23	-0.274087	d_q13c_Never
24	-0.0359449	j_q04a_Yes
25	0.162247	edcat8_Post-secondary, non-tertiary (ISCED 4A-B-C)
26	0.252905	edcat8_Tertiary – bachelor degree (ISCED 5A)
27	0.157216	edcat8_Tertiary – master degree (ISCED 5A)

	Coefficients	Feature
28	0.0540491	edcat8_Tertiary – professional degree (ISCED 5B)
29	-0.0268396	edcat8_Tertiary – research degree (ISCED 6)
30	0.15142	g_q07_Yes
31	0.0613931	g_q05d_Every day
32	-0.0462954	g_q05d_Less than once a month
33	0.0218131	g_q05d_Less than once a week but at least once a month
34	0.0188032	d_q14_Extremely satisfied
35	0.0423613	g_q05a_Every day
36	0.0798542	g_q05e_Every day
37	0.0237213	g_q05e_Never
38	0.121532	g_q05f_Every day
39	0.0351061	g_q05g_Less than once a month
40	-0.00863865	g_q05g_Never
41	0.0140042	g_q05h_Every day
42	-0.352158	g_q05h_Never
43	0.259307	f_q07b_Yes
44	0.107441	b_q10c_Somewhat useful
45	-0.056507	d_q12c_7 to 11 months
46	-0.141571	d_q12c_Less than 1 month
47	0.0394685	d_q12c_None
48	0.141345	g_q05c_Every day
49	-0.137027	g_q05c_Never
50	-0.0905359	d_q12b_A lower level would be sufficient
51	-0.138258	d_q12b_This level is necessary
52	0.00834744	d_q06a_11 to 50 people
53	0.145794	d_q06a_251 to 1000 people
54	0.106445	d_q06a_51 to 250 people
55	0.162631	d_q06a_More than 1000 people
56	-0.238538	d_q03_The private sector (for example a company)
57	0.56252	g_q04_Yes
58	0.78252	b_q10a_Yes
59	0.0366449	g_q06_Moderate
60	0.11028	skill_1

References

Acemoglu, D., Autor, D.H., 2011. Skills, tasks and technologies: Implications for employment and earnings, in: Handbook of Labor Economics. Elsevier, pp. 1043–1171.

Acemoglu, D., Pischke, J.-S., 1999. Beyond becker: Training in imperfect labour markets. The economic journal 109, 112–142.

Autor, D.H., Dorn, D., 2013. The growth of low-skill service jobs and the polarization of the us labor market. American Economic Review 103, 1553–97.

Becker, G.S., 1962. Investment in human capital: A theoretical analysis. Journal of political economy 70, 9–49.

De La Rica, S., Gortazar, L., Lewandowski, P., 2020. Job tasks and wages in developed countries: Evidence from piaac. Labour Economics 65, 101845.

GESIS - Leibniz Institute for the Social Sciences, n.d. Programme for the international assessment of adult competencies (piaac).

Konings, J., Vanormelingen, S., 2015. The impact of training on productivity and wages: Firm-level evidence. Review of Economics and Statistics 97, 485–497.

Lynch, L.M., 1992. Private-sector training and the earnings of young workers. The American Economic Review 82, 299–312.

Lynch, L.M., 1991. The role of off-the-job vs. On-the-job training for the mobility of women workers. The American Economic Review 81, 151–156.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 267–288.