Utrecht University

Importance of Employee Representatives on Achieving Profit Goal for European Companies Using Random Forest

Master's Thesis: Applied Data Science
INFOMTADS

Yan Tan
4698673

Project Supervisor: Dr. Yolanda Grift
Second Reader: Dr. Annette van den Berg

June, 2022

Contents

Abstract

This study explores the importance of both works councils and trade unions on corporate performance using data from the European Company Survey 2019. In total 25 countries are included and each country is fitted with a random forest model. The parameters of national level models are chosen by the out-of-bag error from the whole dataset. Furthermore, the importance of each predictors are compared by Mean Decrease Accuracy. The outcome has been divided into 3 classes: companies who matched, outperformed or underperformed their profit expectation at the end of 2018. As the dataset is highly skewed, oversampling is applied to balance the proportion of different classes.

Random forest does not show a strong performance in this study. The model accuracy ranges between 82-85%. Considering the majority outcome class takes around 82% of the whole observations, it indicates that random forest classification is not reliable for this dataset. The results show that in general, employee representatives are not the most decisive factors for matching the profit goal. Skills of employees, changes in company products or service, and human resource management have more influence on predicting those minorities who beat or fail their profit goals.

Keywords: employee representation, random forest, firm performance, ECS2019

1. Introduction

In 1848, John Mill addressed that the source of "labour's disadvantage" is that individual workers do not have sufficient information to assess their wages and position in the market. Formal associations of workers can help gather information and interpreted labour service into true market worth. The history of labour unions in Europe can be dated back to medieval guilds. Nowadays, employee representations aim at negotiating with management on employee's rights on wages, working hours and conditions, benefits and other issues (EuroWork, 2020). Since 1975, the EU had issued several directives to safeguard the employees' rights at the company level. Two EU directives, Council Directive 75/129 (1975) and Council Directive 77/187 (1977), were the first to make collective employee representation mandatory at the company level. Over time, the requirement of employee representations has deepened and expanded. Representatives, including trade unions and works councils, have played an important role in the European social model.

The relation between employee representatives and the financial performance of the firm has been studied extensively. In 2004-2005, the European Company Survey (ECS) which focused on work conditions amongst establishments in 21 European countries has been carried out. Over years, this survey has been modified three times and expanded its coverage to over 20,000 establishments. For its advantage of the large scale and cross-national coverage, ECS has been studied by economists and sociologists multiple times before. However, analysis with supervised learning, to our knowledge, has yet hardly been done.

In this study, we would like to explore ECS data with supervised learning methods and mainly focus on the importance of employee representations to the corporation performance. The main method for analysis here is random forest classification. To validate the accuracy of the model, the dataset is split into training and test sets. The importance of each predictor is ranked and used as the criteria for comparison. The aim of this paper is to test if works councils and unions are important for companies in forecasting the profit.

2. Literature review

2.1 Influence of worker representatives on the establishment performance

Adam Smith was believed to be the first economist who addressed the problems that labourers have with management and its effect on the economy. His *Wealth of Nations* spawned the systematic study of labour organizations. As has been pointed out, both labour organizations and employers can form cartels to strengthen their own power or to offset the power from the other side (Smith, 1776). Economy wise, two main employee representatives in Europe, trade unions and works councils, seem to impact the establishment performance in opposite ways.

As the main object of trade union is to help workers to maintain or improve the work conditions, many studies revealed the negative impact of unions on the firm performance. With the sample of 979 samples of Japanese manufacturing companies, Brunello (1992) found that unions significantly reduced both productivity and profitability. Regular wages were also reduced in the union firms. The effects are more notable for big firms than small or medium-size ones. With the firm-level variables being controlled, an establishment is more likely to have a strike in a country with more confederations, high membership rates, and a fragmented union movement (Jansen, 2014). Furthermore, the hold-up problem arise when union refrain from cooperating with the management due to the concerns of losing bargaining power. This means while union power being increased, shareholders' profits are being decreased. Investors may also invest less than planned when realising this issue (Grout, 1984).

On the other hand, even though works councils do not increase payment or job satisfaction for the employees (Grund and Schmitt, 2011), they are believed to cooperate with firms and help with the performance. As the firm-level representative of the workforce, works councils in general cannot bargain over wages or work conditions. They are mostly functioning to foster the communication between management and workers and to negotiate issues that are not covered by collective agreements with unions. Freeman and Lazear (1995) offered a model of works council which addressed that works councils can reduce economic inefficiencies by negotiating the worker demands with the firm. With consultation rights, councils can also provide new solutions to problems and smooth the communication between workers and management. The job security can also be increased as employees participate more in the firm decisions. Employees then tend to invest more in skills and give concessions in certain occasions. Addison et al. (2000) examined Freeman-Lazear model later with the data from Germany. The result shows that for large German establishments, the mandatory works councils do not hurt, sometimes even boost the financial performance. This finding is consistent with what Freeman and Lazear (1995) suggested before.

This study will further explore the impact of both works council and trade unions on the establishment performance, specifically whether the establishment can reach its profit goal or not.

2.2 Random Forest

Decision trees have been widely used in classification mining. The non-parametric prediction method has brought a lot of advantages including no assumption needed for the training sample and ability of handling incomplete and qualitative data (Joos et al., 1998). By averaging over a great amount of decision trees, random forest manages to have a low variance while maintaining the low bias. Besides the good performance in forecasting (Breiman, 1996), it also reports the information about the importance of each individual predictors. This method has been applied to various fields. In 2019, Behr, Schiwy and Weinblat applied random forest for default prediction in 7 EU countries. It shows that random forest does not only provide the good prediction performance, but can also give out the local-level peculiarities for horizontal comparison. In another study Weinblat did in 2018, random forest was applied for forecasting European high-growth firms (Weinblat, 2018). It indicates the great performance of random forest in determining relevant predictors. Among machine learning methods, Patel et al. (2015) compared the performance of support vector machines, artificial neural networks and random forests on predicting direction of stock movements in India. Random forests outperformed other methods on overall performance. Given the considerable number of works showing the optimistic result for random forest, this study would also adopt random forest in the tests.

3. Data and Methods

In this study, the data was extracted from the ECS 2019. It was collected by Eurofound and covers topics including work-life balance, flexible contracts, workplace innovation, and social dialogue. Corporations from 28 countries, including the United Kingdom and 27 EU Member

States, participated in this survey. Two questionnaires are included: Employee Representative Questionnaire (ER) and Management Questionnaire (MM). For the analysis in this study, only MM data is used.

3.1 Data wrangling

Considering the large amount of variables in the data, only 20 has been chosen for this study as the corresponding variables that Van den Berg, A., et al. (2013) applied for the 2009 ECS survey. For questions being asked in 2009 but not in 2019, we omitted the variables. Table 1 lists all variables used in this study.

Original variable's name in ECS2019	**Question in ECS2019**
countrycode	2-digit ISO code
mmerconfirm_v3_3	Works council - official employee representation exist
mmerconfirm_v4_3	Works council - official employee representation doesn't exist
mmerconfirm_v3_1	Trade union representation - official employee representation exist
mmerconfirm_v4_1	Trade union representation - official employee representation currently doesn't exist
skillsmatch_d	What percentage of employees have the skills that are about right to do the job?
overskill_d	What percentage of employees have a higher level of skills than is needed in their job?
vpbres_d	Payment by results, for example piece rates, provisions, brokerages - employees received variable pay
vpinper_d	Extra pay linked to individual performance - employees received variable pay
vpgrpe_d	Extra pay linked to the performance of the team, working group - employees received variable pay
vpprsh_d	Extra pay linked to the results of the company - employees received variable pay
emporg	Is the company a member of any employers' organisation which participates in the negotiation of collective agreements?
est_size	Establishment size in number of employees
mm_sector_grp2	Collapsed sector group variable
sickleave	Do you think the level of sickness leave in this establishment is too high?

retainemp	How difficult is it for this establishment to retain employees?
chempfut	In the next three years, how do you expect the total number of employees in this establishment to change?
prodvol	Since 2016, how has the amount of goods or services produced by this establishment changed?
profit	In 2018, did this establishment make a profit?
profplan	Did this establishment expect to make a profit in 2018?

Table 1: selected variables

Source: ECS 2019

As companies either answered to the existence of the employee representatives (variable mmerconfirm_v3_3/ mmerconfirm_v3_1) or the non-existence of it (variable mmerconfirm_v4_3/ mmerconfirm_v4_1), these questions were merged into variables “WCs” and “Unions” to get rid of the massive amount of missing values. If the company did not answer either question, it was labelled as “skipped”. Since the dependent variable being studied here is whether the establishments have reached their expectation of profit, non-profit companies and those who skipped the questions were not considered in this study. Variables “profit” and “profplan” were merged into one new variable “performance” which contains 3 classes: actual performance exceeded, matched or failed the expectation.

Of the 27 EU member states, three countries - Cyprus, Malta, and Sweden – did not answer questions related to works councils or trade unions. Despite the fact that EU directives have prompted new arrangements for non-unionized employees, in Malta, it is still mainly the union that represents the employees at work. In Sweden and Cyprus, only trade unions provide employee representation. There is no other elected structure as workers councils (Fulton, 2021). Given the primary interest in the influence of both works council and trade unions, these three countries had been excluded from further analysis.

3.2 Random Forest

3.2.1. Random forest

Random forest is an ensemble classifier which consists of a collection of decision trees. It was first introduced by Leo Breiman in 2001. This approach can be used for both regression and classification. In this study, it was applied as a classification method. To get random forest from a single decision tree, the first step is to generate new data from the original dataset. Each records can be selected multiple times into the new data. Eventually the generated data should have the same amount of records as the original one. Every new generated data is one “tree” that the random forest has. This process of random sampling with replacement is called “bootstrap”.

The randomness in this method is not only shown in the bootstrap, but also in random variable selection. At each decision tree node (where the tree splits), random features are selected to determine the split. For any record, each tree gives it an outcome class. In the end, there will be an aggregation where random forest would cast a vote for the most popular class from the all the results.

With a great amount of decision trees generated randomly, the result is more accurate and stable than individual decision trees. However, it is very difficult to visualize random forest. In the case of a random forest with 500 trees, theoretically all 500 trees can be drawn out. At the meantime, for the randomness mentioned above, two parameters need to be tuned for higher accuracy: the number of trees and the number of features in each split.

3.2.2. The model

The main object in this study is to test the influence of the works council and trade union on reaching the profit goals for companies. There are in total 3 classes in the outcome which are the ones whose actual performance exceeded, matched or failed the expectation. The majority had matched their expectation with the proportion of 82%. This high percentage will cause the problem of imbalanced classification which makes minority classes less well modelled and lures the algorithm to classify most observations as the majority class (Fernández et al., 2017). To address this issue, a common approach is resampling. It can be done by either undersampling (to remove the data from the majority) or oversampling (to add more examples to the minority). Since undersampling may cause information loss, we balanced the weights of each class by oversampling. The minority classes have been duplicated one time and three times respectively (Table 2). The number of duplication is chosen due to the consideration of both reducing the proportion of majority class and keeping data volume within computing capacity of R.

Performance	Amount before oversampling	Proportion before oversampling	Amount before oversampling	Proportion after oversampling
Failed the expectation	2252	11,94%	4504	18,54%
Matched the expectation	15551	82,44%	15551	64,01%
Exceeded the expectation	1060	5,62%	4240	17,45%

Table 2: different outcome classes before and after oversampling

Source: ECS 2019

80% of the observations have been separated as the training set and 20% as the test set. Parameters (number of trees and number of features) were tuned as a whole dataset in one random forest. Next, the assumption has been made that if the parameters suit this whole dataset, they should also work for the models of different nations. Out-of-bag (OOB) error is used for choosing the optimal parameters. OOB error rate is the misclassification rate of the leftover data (default 1/3 of the whole data) from each tree. Figure 1 indicates that the OOB error reaches the bottom when the number of predictors is 4. Hence 4 is chosen to be the number of predictors. The error rate improves significantly as the number of trees grows to 50, then stays flat afterward (Figure 2). As default number of trees is normally 500 and more trees would generally give higher accuracy, the number of trees for all the nation-wise models remains 500.

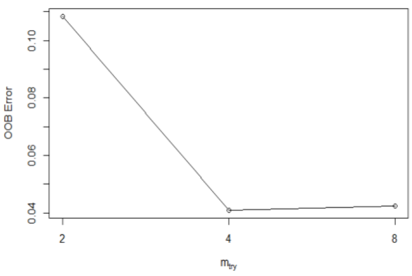

Figure 1: OOB error with different numbers of predictors (mtry)

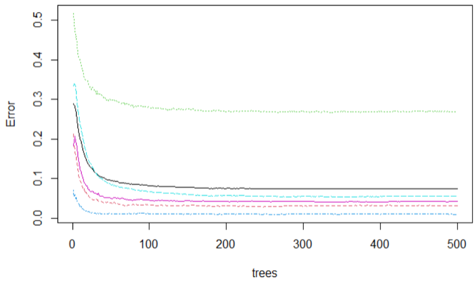

Figure 2: Error with different numbers of trees[1]

Source: ECS 2019

After removing Sweden, Cyprus, and Malta from the dataset, in total of 25 countries were left. For each country, a random forest model was fitted while the importance of variables were compared among nations. The importance

[1] Figure 2: Error for different classes (coloured) and out of bag samples (black)

can be illustrated by two criteria, Mean Decrease Accuracy and Mean Decrease Gini. These two measurements normally show very high similarities. The higher the value of mean decrease accuracy or mean decrease Gini have, the more important the variable is in the model. Mean Decrease Gini, which is more related to the local decision function, measures the importance of the variable on the change of Gini impurity (Formula 1) at each split. As we would like to target on the overall model performance, Mean Decrease Accuracy would be the main measurement. Mean Decrease Accuracy shows how less accurate the model would be if one variable being dropped out. It can be further broken down to different outcome classes.

$$Gini\ Impurity = \sum_{l=1}^{L} \hat{p}_{ij}(1 - \hat{p}_{ij})$$

Formula 1[2]

$$Mean\ Decrease\ Accuracy = \frac{Mean(Decreases\ in\ Accuracy\ of\ Trees)}{StandardDeviation(Decreases\ in\ Accuracy\ of\ Trees)}$$

Formula 2

4. Results

One point to note here is as random forest generates the trees and features completely randomly, the result slightly varies each time. However, considering the large amount of trees being formed, the difference is very subtle. After multiple times of testing, the accuracy of the test set stays within 82%-85%. Considering our majority class takes 82% of the share, this model is not much better than classifying everything into the majority class. In summary, random forest might not be the best model here for the ECS 2019 dataset.

To compare the importance of the same variables in different models, we rank the variables by their Mean Decrease Accuracy. Table 3 lists the variables that have ever been ranked within the top three in any nation. The most important variable is the percentage of employees with matching skills in the job. Among 25 countries, this variable showed with top 3 highest importance for 21 times. Other variables related to human resource management, including the different formats of payment, are also showing significant importance.

[2] Formula 1: pij represents the proportion of training observations in the ith region that are from the jth class

Variables	Description	count
skillsmatch_d	What percentage of employees have the skills that are about right to do the job?	21
prodvol	Since 2016, how has the amount of goods or services produced by this establishment changed?	14
overskill_d	What percentage of employees have a higher level of skills than is needed in their job?	12
vpbres_d	Payment by results, for example piece rates, provisions, brokerages - employees received variable pay	11
vpinper_d	Extra pay linked to individual performance - employees received variable pay	10
vpprsh_d	Extra pay linked to the results of the company - employees received variable pay	4
vpgrpe_d	Extra pay linked to the performance of the team, working group - employees received variable pay	3

Table 3:Variables with top 3 importance in any nation

Source: ECS 2019

The importance rankings of works council and trade union in each nation are listed in Table 5 and 6. Table 4 indicates that works councils and trade unions are not generally showing strong influence in the model. Despite that trade unions seem to have a better impact on reaching the profit goal in several countries, the average ranking of trade unions and works councils among 17 independent variables are both around 13.

Country	**AT**	**BE**	**BG**	**CZ**	**DE**	**DK**	**EE**	**EL**	**ES**	**FI**	**FR**	**HR**	**HU**	**IE**	**IT**	**LT**	**LU**	**LV**	**NL**	**PL**	**PT**	**RO**	**SI**	**SK**	**UK**
Union	12	12	11	15	15	9	10	15	11	14	14	7	13	15	14	11	15	15	9	14	15	11	15	9	14
WCs	15	10	10	11	12	12	13	12	15	11	10	14	15	13	15	12	14	14	15	12	14	14	11	11	15

Table 4:Variables importance ranking (overall Mean Decrease Accuracy) - Trade Union and Works Council

Source: ECS 2019

However, when we break the Mean Decrease Accuracy on outcome class, one notable finding is that compared to other classes, the employee representatives have a significantly higher importance in the class in which final performance matched their expectations. Table 7 and Table 8 illustrate the big gap between the matching class and other classes. The largest difference is provided by Portugal. Its importance ranking of trade union improved from No.15 to No.5 when we switched from the overall Mean Decrease Accuracy to the matching class only. Within countries of Austria, Luxembourg, Hungary, Greece, France, Latvia, and Portugal, the difference is remarkable for both types of employee representatives. Meanwhile, there are still few countries - including Ireland, Italy, Lithuania and Romania - that do not show much overall importance of employee representatives and stay consistent among all the outcome classes.

Country	Failing	Matching	Surpassing	Overall MDA
AT	15	6	15	15
BE	13	6	11	10
BG	10	15	9	10
CZ	11	2	12	11
DE	14	10	14	12
DK	14	8	11	12
EE	13	12	12	13
EL	14	5	12	12
ES	14	8	14	15
FI	13	15	12	11
FR	11	3	10	10
HR	14	6	14	14
HU	15	9	14	15
IE	14	14	13	13
IT	15	14	15	15
LT	10	10	12	12
LU	15	8	14	14
LV	14	6	14	14
NL	15	12	15	15
PL	12	15	12	12
PT	14	6	14	14
RO	13	13	12	14
SI	13	3	10	11
SK	9	7	12	11
UK	15	15	14	15

Table 5: Variables importance ranking broken down on outcome class - Works Council

Source: ECS 2019

Country	Failing	Matching	Surpassing	Overall MDA
AT	14	5	12	12
BE	14	10	12	12
BG	13	10	12	11
CZ	15	15	15	15
DE	15	11	15	15
DK	9	7	8	9
EE	12	3	10	10
EL	15	9	15	15
ES	11	9	11	11
FI	11	8	14	14
FR	14	7	15	14
HR	7	4	7	7
HU	14	7	13	13
IE	13	13	15	15
IT	14	15	13	14
LT	12	11	11	11
LU	14	11	15	15
LV	15	9	15	15
NL	10	1	8	9
PL	14	13	14	14
PT	15	5	15	15
RO	12	11	11	11
SK	15	12	14	15
UK	12	11	9	9

Table 6: Variables importance ranking broken down on outcome class - Trade Union

Source: ECS 2019

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
AT						○									▲
BE						○				▲					
BG										▲					○
CZ		○									▲				
DE										○		▲			
DK								○				▲			
EE												○	▲		
EL					○							▲			
ES								○							▲
FI											▲				○
FR			○							▲					
HR						○								▲	
HU									○						▲
IE													▲	○	
IT														○	▲
LT										○		▲			
LU								○						▲	
LV						○								▲	
NL												○			▲
PL												▲			○
PT						○								▲	
RO													○	▲	
SI			○								▲				
SK							○				▲				
UK															○/▲

○： MDA of the matching class

▲： overall MDA

Table 7: Difference in variables importance ranking - Works Council

Source: ECS 2019

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
AT					○							▲			
BE										○		▲			
BG										○	▲				
CZ															○/▲
DE											○				▲
DK							○		▲						
EE			○							▲					
EL									○						▲
ES									○		▲				
FI								○						▲	
FR							○							▲	
HR				○			▲								
HU							○						▲		
IE													○		▲
IT											○/▲				
LT											▲				
LU											○				▲
LV									○						▲
NL	○								▲						
PL													○	▲	
PT					○										▲
RO											○/▲				
SI												○			▲
SK									▲		○				
UK														○/▲	

○： MDA of the matching class
▲： overall MDA

Table 8: Difference in variables importance ranking - Trade Union

Source: ECS 2019

Even though the matching skills of employees shows overall the highest importance, Table 9 indicates that for most countries, when looking at matching class only, the importance of skills dropped slightly. For few like Estonia, Greece and France, the difference seems to be more significant. However, for Estonia and France, another variable “overskill_d”, which is also related to the employees’ skills, takes the top 3 importance instead.

Country	Failing	Matching	Surpassing	Overall MDA
AT	1	4	2	2
BE	1	5	1	1
BG	2	9	6	4
CZ	3	6	2	2
DE	1	1	2	2
DK	1	1	1	1
EE	4	13	5	5
EL	2	10	2	2
ES	2	5	2	2
FI	1	3	3	3
FR	4	10	3	1
HR	1	7	1	1
HU	3	5	3	3
IE	1	3	1	1
IT	2	4	1	1
LT	5	2	6	5
LU	2	1	2	2
LV	1	5	3	2
NL	1	4	1	1
PL	3	4	3	3
PT	2	4	2	2
RO	2	7	1	2
SK	4	5	6	4
UK	5	3	3	3

Table 9: Variables importance ranking broken down on outcome class – matching skills

Source: ECS 2019

Considering the majority here is the matching class, this result illustrates that generally both trade union and works council have a certain impact on deciding if a company can reach its profit goal. For those who outperform or underperform their expectation, the cause might be some action related to human resource management or an unusual proportion of skilled employees. Although this is not within the scope of this paper, further research could be valuable in this area.

5. Discussion

5.1 Dataset limitation

The limitation of the ECS dataset might lead to certain bias in the result. This survey was answered by one management respondent for each company, which would make subjective answers unavoidable, especially for the variable “sickleave” (if the employer find the level of sickness leave too high) and “retainemp” (how difficult it is to retain employees).

As most questions from the survey are multiple-choice questions, the majority of variables are character (answers in different classes) instead of numeric (for example, the amount of the actual profit). This might cause the loss of complexity and precision in the questions. There are only two questions in direct relation to the company outcomes: if the establishment was expected to make a profit and if it actually made a profit in 2018. Over 80% of the companies answered yes to the expectation of profit while eventually, around 72% made a profit and 10% broke even. The measurement of outcome has been simplified into two yes-and-no questions which neglect the comprehensiveness of finance.

5.2 Data wrangling

Random oversampling was applied in this study to balance class distribution. The method being used in this study is naive oversampling, which is to duplicate the minority classes without modification. There are several other different methods for oversampling. One of the most popular techniques is SMOTE: Synthetic Minority Over-sampling Technique. SMOTE takes a sample of the minority class from the dataset. New synthetic data will be randomly generated between the sample and its k nearest neighbours (Chawla et al., 2002). Augmentation can also oversample the dataset by adding slightly modified minority copies. These methods can be tested in the future for the improvement of model accuracy.

5.3 Random Forest

In this study, we used the importance ranking among all variables to compare the influence of employee representatives in different nations. This is due to the consideration that even with the same parameters in random forest models, different countries show different sensitivities. However, it may cause other problems. For countries having all variables with similar importance, rankings might enlarge the difference than it actually is.

5.4 Future study

From the 15,551 observation of matching class in Table 2, around 7% (1,086 companies) had no expectation of making profit and eventually did not make any. These observations has been grouped together with 14,465 observations who expected themselves to make profit and matched their expectation. Despite the proportion is relatively small, the result can still be different if the matching class can be further split into with and without profit.

In total 20 out of 385 questions were selected from the survey. As this study intended to have variables with more general applicability, certain types of questions haven’t been considered in the model. For example, changes in the establishment. More features can be tested in the future for model modification.

5.5 Conclusion

This study investigated the question of how important employee representatives is to a company’s profit. In total 25 European countries are included. Each nation is fitted with a random forest model. The importance of variables were compared by Mean Decrease Accuracy.

The accuracy of random forest model is ranging from 82-85%, which states that using random forest for the ECS 2019 dataset is not reliable. This might be caused by the limitation of the data type and the skewed distribution of outcome classes. Among the three outcome classes, 82% companies matched their profit expectation while 12% outperformed and 6% underperformed their goal. Results show that, on the one hand, the skill level of employees is significantly important for forecasting the outcome. On the other hand, the importance of employee representatives grows much higher when we only look at the companies that reached their profit goals. This means that, as the majority of companies can reach the expectation at the end of the year, trade unions and works councils do show certain importance. However, matching skills of employees, human resource management and changes in the provided product or service have more influence on deciding if the establishment is going to surpass or fail its profit goal.

6. References

Addison, J., Siebert, S., Wagner, J. and Wei, X., 2000. Worker Participation and Firm Performance: Evidence from Germany and Britain. *British Journal of Industrial Relations*, 38(1), pp. 7-48.

Behr, A., Schiwy, C. and Weinblat, J., 2019. Investment, default propensity score and cash flow sensitivity in six EU member states: evidence based on firm-level panel data. *Applied Economics*, 51(49), pp. 5345-5368.

Breiman, L., 1996. Random Forests. *Machine Learning* 45, pp. 5–32.

Brunello, G., 1992. The Effect of Unions on Firm Performance in Japanese Manufacturing. *Industrial and Labor Relations Review*, 45(3), pp. 471–487.

Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, pp. 321-357.

Council Directive 75/129/EEC of 17 February 1975 on the approximation of the laws of the Member States relating to collective redundancies. *Official Journal* L 048, 22/02/1975, pp. 29-30.

Council Directive 77/187/EEC of 14 February 1977 on the approximation of the laws of the Member States relating to the safeguarding of employees' rights in the event of transfers of undertakings, businesses or parts of businesses. *Official Journal* L 061, 05/03/1977, pp. 26-28.

EuroWork. 2020. Employee representation. [online] Available at: <https://www.eurofound.europa.eu/observatories/eurwork/industrial-relations-dictionary/employee-representation> [Accessed 26 June 2022].

Fernández, A., del Río, S., Chawla, N. and Herrera, F., 2017. An insight into imbalanced Big Data classification: outcomes and challenges. *Complex & Intelligent Systems*, 3(2), pp. 105-120.

Freeman R. B. and Lazear E. P., 1995. An Economic Analysis of Works Councils. *National Bureau of Economic Research*, pp. 27-52.

Fulton, L. 2021, National Industrial Relations, an update (2019-2021). *Labour Research Department and ETUI*. [online] De.worker-participation.eu. Available at: <http://www.worker-participation.eu/National-Industrial-Relations>[Accessed 26 June 2022].

Grout, P., 1984. Investment and Wages in the Absence of Binding Contracts: A Nash Bargaining Approach. *Econometrica*, 52(2), pp. 449-460.

Grund, C. and Schmitt, A., 2011. Works Councils, Wages, and Job Satisfaction. *SSRN Electronic Journal*, 45(3), pp. 299-310.

Jansen, G., 2014. Effects of Union Organization on Strike Incidence in EU Companies. *ILR Review*, 67(1), pp. 60-85.

Joos, P.P.M., Vanhoof, K., Ooghe, H. & Sierens, N., 1998, Credit classification: A comparison of logit models and decision trees. *Applications of machine learning and data mining in finance*. 10th European Conference on Machine Learning, pp. 59-72.

Mill, J., 1848. *Principles of political economy*. Reprint, Fairfield, NJ: Kelley, 1987.

Patel, J., Shah, S., Thakkar, P. and Kotecha, K., 2015. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. *Expert Systems with Applications*, 42(1), pp. 259-268.

Smith, A., 1776. *Wealth of Nations*, Book 1, Ch. 8. Wordsworth Classics of World Literature, Hertfordshire: Wordsworth Editions Limited.

Van den Berg, A., Grift, Y., Witteloostuijn, A. van, Boone, C., and Brempt, O. Van der, 2013. The effect of employee workplace representation on firm performance. A cross-country comparison within Europe. *Utrecht School of Economics, Tjalling C. Koopmans Research Institute*, 13(05).

Weinblat, J., 2018. Forecasting European high-growth Firms - A Random Forest Approach. *Journal of Industry, Competition and Trade*, 18(3), pp. 253–294.

Appendix

R code

~~~
---
title: "Thesis"
author: "Yan"
date: '2022-05-04'
output: html_document
---

Dataset:

http://nesstar.ukdataservice.ac.uk/webview/index.jsp?v=2&mode=documentation
&submode=abstract&study=http://nesstar.ukdataservice.ac.uk:80/obj/fStudy/86
91&top=yes


```{r}
library(foreign) #to read dta
library(dplyr)
#below for Random Forest
library(randomForest)
#below for confusion Matrix
library(caret)

```

```{r}
ecs2019_mm<-read.csv("U:/Thesis/data/2019/csv/ecs2019_mm_ukds.csv")
#ecs2019_mm<-
read.csv("C:/Users/silvi/OneDrive/Desktop/thesis/data/ecs2019_mm_ukds.csv")
```

select variables we need
```{r}
select_2019<-subset(ecs2019_mm,select = c(countrycode,mmerconfirm_v3_3,
mmerconfirm_v4_3, mmerconfirm_v3_1, mmerconfirm_v4_1, skillsmatch_d,
overskill_d,vpbres_d, vpinper_d, vpgrpe_d,
vpprsh_d,emporg,est_size,mm_sector_grp2,sickleave,retainemp,
chempfut,prodvol,profit, profplan))
```

```{r}
#structure of the data
str(select_2019)
```

```{r}
summary(select_2019)
```

```{r}
#deal with missing values
mutate_2019<-select_2019
mutate_2019$mmerconfirm_v3_3[is.na(mutate_2019$mmerconfirm_v3_3)] <- 0
mutate_2019$mmerconfirm_v3_1[is.na(mutate_2019$mmerconfirm_v3_1)] <- 0
mutate_2019$mmerconfirm_v4_3[is.na(mutate_2019$mmerconfirm_v4_3)] <- 0
mutate_2019$mmerconfirm_v4_1[is.na(mutate_2019$mmerconfirm_v4_1)] <- 0
~~~

~~~~
mutate_2019$profplan[is.na(mutate_2019$profplan)] <- -3
```

```{r}
#reduce the no. of variables
mutate_2019<-mutate_2019%>%mutate(

WCs=ifelse(mmerconfirm_v3_3==1,"yes",ifelse(mmerconfirm_v4_3==2,"yes",ifels
e(mmerconfirm_v4_3==0&mmerconfirm_v3_3==0,"skipped","no"))),

Unions=ifelse(mmerconfirm_v3_1==1,"yes",ifelse(mmerconfirm_v4_1==2,"yes",if
else(mmerconfirm_v4_1==0&mmerconfirm_v3_1==0,"skipped","no"))),
  performance=ifelse(profit==-7,as.integer(-3),ifelse(profplan==-
3,as.integer(0),ifelse(profplan==profit,as.integer(1),ifelse(profplan!=prof
it&profplan==1,as.integer(-2),as.integer(2))))))
```
Sweden has no information of WCs or Unions. all companies (in total 122)
from CY and MT (in total 145) did not answer these questions.

```{r}
#remove SE/MT/CY as there is no WCs info
mutate_2019<-
filter(mutate_2019,countrycode!="SE"&countrycode!="MT"&countrycode!="CY")
```

```{r}
mutate_2019<-filter(mutate_2019,performance!="-3"&performance!="0")
```

```{r}
mutate_2019<-mutate_2019 %>%
  mutate(across(where(is.integer), factor)) #change each column into factor
mutate_2019$skillsmatch_d<- as.integer(mutate_2019$skillsmatch_d) #change
back to integer
mutate_2019$overskill_d<- as.integer(mutate_2019$overskill_d)
```

```{r}
#get rid of four columns
mutate_2019<-subset(mutate_2019,select = c(-mmerconfirm_v3_3, -
mmerconfirm_v4_3, -mmerconfirm_v3_1, -mmerconfirm_v4_1,-profit,-profplan))
```

```{r}
#split into train and test
set.seed(123)
ind <- sample(2, nrow(mutate_2019), replace = TRUE, prob = c(0.8, 0.2))
train_2019_bf <- mutate_2019[ind==1,]
test_2019 <- mutate_2019[ind==2,]
```

```{r}
train_2019_bf %>% count(performance)
```

```{r}

#for class -2
class_neg2<-filter(train_2019_bf,performance=="-2")
rows= c(1:nrow(class_neg2))
times = 1
~~~~

~~~
class_neg2<-class_neg2[rep(rows, times),]


#for class 2
class_2<-filter(train_2019_bf,performance=="2")
rows= c(1:nrow(class_2))
times = 3
class_2<-class_2[rep(rows, times),]

train_2019<- rbind(class_neg2, class_2,train_2019_bf)
train_2019 %>% count(performance)
```

```{r}
# For reproducibility
set.seed(222)

# Machine Learning: Random Forests
# Default forests grows 500 trees!
# Trees: ntree, default 500
# Variables, mtry, default = sqrt(p) for classification and p/3 for
regression
# p is number of features
# Draw ntree bootstrap samples
# mtry predictors at each node
rf_bf <- randomForest(performance~., data=train_2019)
```

```{r}
# Prints out the confusion matrix!
# mtry sqrt(21) is approx 4
print(rf_bf)
```

```{r}
# Allows us to look at the different attributes we can take
# We can drag out confusion matrix with rf$confusion
attributes(rf_bf)
```

```{r}
# Prediction & Confusion Matrix - train data
# Training Accuracy: 97.9%

p1 <- predict(rf_bf, train_2019)
confusionMatrix(p1, train_2019$performance)
```

```{r}
# Prediction & Confusion Matrix - test data
# Testing Accuracy: 76.2%
p2 <- predict(rf_bf, test_2019)
confusionMatrix(p2, test_2019$performance)
```

```{r}
# Error rate of Random Forest
# We observe that as the number of trees grow, error rate doesn't seem to
improve
# Error for your different classes (colored) and out of bag samples (black)
plot(rf_bf)
~~~

~~~~
```
▯ Higher the value of mean decrease accuracy or mean decrease gini score ,
higher the importance of the variable in the model.
```{r}
importance(rf_bf)
```
```{r}
varImpPlot(rf_bf,
           sort = T, #Should the variables be sorted in decreasing order of
importance?
           n.var = 10, #How many variables to show?
           main = "Top 10 - Variable Importance" #plot title
           )
```

```{r}
t <- tuneRF(train_2019[,-17], train_2019[,17], #column forth to last is our
predictor variables
  stepFactor = 0.5, #  at each iteration, mtry is inflated (or deflated) by
this value
  plot = TRUE, #whether to plot the OOB error as function of mtry
  ntreeTry = 200, # number of trees used at the tuning step
  trace = TRUE, #whether to print the progress of the search
  improve = 0.05 #the (relative) improvement in OOB error must be by this
much for the search to continue
  )
```
Two parameters are important in the random forest algorithm:
Number of trees used in the forest (ntree ) and
Number of random variables used in each tree (mtry ).


```{r}
# Prediction & Confusion Matrix - test data

#p2 <- predict(rf_af, test_2019)
#confusionMatrix(p2, test_2019$performance)
```

seperate nations
```{r}
Cty_combine<-data.frame(matrix(ncol = 8,nrow=0))
colnames(Cty_combine)<-
c("cty_code","X.3","X.2","X0","X1","X2","MeanDecreaseAccuracy","MeanDecreas
eGini"   )
```

```{r}
cty<-unique(mutate_2019$countrycode)
for (i in cty){
  cty_temp<-filter(mutate_2019,countrycode==i)
  rf_temp <- randomForest(performance~., data=cty_temp, ntree = 500, mtry =
8, importance = TRUE,proximity = TRUE)
  cty_name<- paste("cty_",i,sep = "")
  a<-data.frame(importance(rf_temp))
  a<-mutate(a,cty_code=i)
  assign(cty_name,a)
}

rm(a)
rm(cty_temp)
~~~~

~~~
rm(rf_temp)
```
~~~