UNIVERSITY OF TARTU
Faculty of Social Sciences
School of Economics and Business Administration

Gratian Allan

PROFILING OF REGISTERED UNEMPLOYED

Master's thesis

Supervisor: Andres Võrk
Co-supervisor: Jaan Masso

Tartu 2019

Name and signature of supervisor……

Allowed for defence on……..

(date)

I have written this master's thesis independently. All viewpoints of other authors, literary sources and data from elsewhere used for writing this paper have been referenced.

…….

(signature of author)

TABLE OF CONTENTS

1. ABSTRACT

Long term unemployment is a grave social issue resulting in sustained repercussions on the society which can be tackled through statistical profiling of the unemployed. Therefore, through this thesis, my endeavour was to test and compare the effectiveness of three existing models - Gradient Boosting machine, Random Forest and Logistic Regression model in predicting long term unemployment using the EUIF (Estonian Unemployment Insurance Fund) database consisting of 256,703 unemployment spells from 2014-2017. The prediction accuracy of these models were compared through their confusion matrices based on certain performance metrics and ROC curves. Although the performance metrics indicate that the overall accuracy of Logistic Regression model is comparable to RF and GBM models, the ability to capture long-term unemployed is better for the latter models. Finally, the results suggest that Random Forest model could be the regarded the best statistical profiling model in predicting long term unemployment as GBM and LR models are vulnerable to overfitting in case of noisy data.

2. INTRODUCTION

Unemployment is a core critical issue in every national economy and has a serious impact on the wellbeing of the population. It can adversely affect not only the unemployed but also the people surrounding them, thereby affecting the whole nation. The unemployment rate of the country is constantly monitored along with GDP and inflation to track the economic health of the nation. Long term unemployment is a grave problem due to several factors such as erosion of human capital, social exclusion, economic and social costs and higher risk of poverty. Since the probability of an unemployed person finding a job decreases over the unemployed spell, it is the paramount responsibility of every national government and concerned regulatory authorities to check for unemployment and ensure that enough opportunities are generated to tackle this systemic issue.

Over the years, several ways to tackle unemployment have been researched upon and implemented in various countries. Some of these means include tax reforms, changes in monetary policies, budget allocations, etc. Although these methods are quite effective in solving unemployment, profiling tools help identify job seekers who are more likely to be long-term unemployed. By doing so, the Public Employment Services (PES) can better serve these individuals by providing services to them more efficiently (Weber, 2011).

Statistical profiling places the jobseekers in different categories based on their need for additional support by PES counsellors (OECD, 2018). It takes into account the time it will take for the individuals to find work based on their skills, capabilities and the vacancy of related jobs in the economy. Profiling helps with providing services on a priority basis to individuals who are at a higher risk of finding employment rather than those who would find work without any additional help from PES, thus increasing the PES' cost efficiency. It can be used to allocate spending based on the results of profiling. The use of statistical profiling is widespread and used extensively in several European countries (Barnes, Wright, 2015). This is possible because of higher availability of real-time data, advancements in computers and computing power, and the use of more complicated statistical models.

While profiling jobseekers, several factors are taken into consideration to optimally conduct the process. Input variables are run through the statistical model to generate an output that identifies jobseekers' probability of becoming long-term unemployed. The input variables include typically age, gender, job search behaviours, pay expectations, skills, education levels, job opportunities available, etc. The generated output categorizes the jobseekers as either low-risk or high-risk of long-term unemployment. The timing of support required for these jobseekers as well as the intensity is determined, and the PES works accordingly to help these individuals.

In recent years, machine learning and artificial intelligence have improved exponentially to be effectively applied in several industries and markets. Statistical profiling with use of artificial intelligence has started to gain some traction. The machine learning algorithms at the one used in Flemish PES in Belgium is one such example (OECD, 2018).

Numerous variables are used to estimate the probability of a jobseeker to remain unemployed for more than six months. Frequent updates to the model ensure that the profiling generated through it is as accurate as possible. An intriguing method used in this model is through 'click data,' which tracks and monitors the activities of jobseekers on the PES website and records the clicks on the different job vacancies listed on there.

The statistical model used in the aforementioned initiative in Belgium uses a Random Forest model. This model, along with the Logistic Regression model and Gradient Boosting model will be discussed in this paper. Logistic Regression does not take into account the possible interactions between the input variables. More modern models such as Random Forest and Gradient Boosting are therefore considered more accurate in predicting the probabilities. Although numerous statistical profiling systems have been abandoned in Europe, this thesis argues that the judgement of individual (Public Employment Services) PES counsellors can be less accurate in predicting the likelihood of long-term unemployment compared to statistical profiling systems. The aim of this thesis is to test if modern statistical ensemble methods, such as Random Forest or gradient bosting machine can outperform simple Logistic Regression models, which has been used in profiling earlier in Estonian studies (Trumm, 2018).

In order to analyze the different models and their accuracies, a thorough study of the different profiling systems implemented by different countries is used to build a knowledge base. The effects of unemployment on the labour market are also taken into consideration. It is found that job trainings, wage subsidy programs, etc. can positively impact the responsiveness of in the event of unemployment. This paper also studies the impact of individual characteristics and region of residence to predict the individual's probability of staying unemployed or long-term unemployed in Estonia (Marksoo,2011).

This thesis also incorporates the findings of Leuvensteijn & Koning (2000) regarding the duration dependence effects, motivation of the unemployed and individual duration effects. Another study by Connell et al (2012) is delved into, to gain an understanding of the several factors affecting the likelihood of remaining unemployed for a year or more. The

findings of a study by Weber (2011) involved the use of a profiling tool which looked into the needs of the jobseekers in terms of risks related to characteristics such as age, gender, etc. According to this study, it is imperative that simple and accurate tools be used instead of more complex tools. The findings provide a strong argument to this thesis, which is aimed at emphasizing the higher accuracy of modern statistical models.

The main objective of this thesis is to test the effectiveness of whether modern statistical ensemble methods, such as Random Forest or Gradient Boosting machine can outperform simple Logistic Regression models. The effectiveness in profiling of long term unemployed earlier in Estonian studies is measured by comparing and analyzing the results derived from the estimations and investigating the predictive strength of the variables. The prediction accuracy of these classification models is then evaluated based on performance metrics such as Accuracy, Precision, Sensitivity, Specificity and False Positive rate. The analytical output and observations are then presented in the form of the generated Receiver Operating Characteristic (ROC) graphs. Furthermore, the evaluation and comparison of the performance of the three models is done by calculating the Area Under the Curve (AUC) of each model. The level of influence of the variables on the Random Forest model and Gradient Boosting model's prediction is explained by the Variable Importance graphs. Accuracy of data is a very important requirement for the reliability of the model output, as evidenced by the research outlined in the literature review in the following section. Therefore, for this analysis, I have used credible data from the EUIF (Estonian Unemployment Insurance Fund) database consisting of 256,703 unemployment spells from 2014-2017.

The next section deals with the theoretical background and review of profiling systems. The third section describes the methodology and statistical framework, followed by the description of the data analysed. The results of the analysis are presented in the fifth section. Finally, the last section concludes the paper.

3. LITERATURE REVIEW

Looking at the new approaches to skills profiling, it is critical to note that factors such as increasing employer demand for transferable skills and competencies, changes in the labour market and frequent job transitions which require greater adaptability have resulted in more complex profiling approaches ,since simply collecting information of an individual job seeker's employment record, work experience and formal qualifications has become insufficient and require further gathering of information on "generic" and "soft" skills.

This review explores the recent trends and innovative approaches of profiling systems while examining the developments in profiling systems along with the progress and challenges for the future. This review particularly emphasizes on the assessment and the use of profiling for risk identification, resource allocation, matching and action planning.

Mroz and Savage (2006) explore the long term impacts of youth unemployment and their labour market outcomes. The study's premise is based on the strong evidence of human catch-up response to unemployment, wherein, an unemployment spell experienced today increases the likelihood that a young person trains in the near future. They state that a dynamic model of human capital accumulation can predict this catch-up response. The authors claim that despite the catch-up response, there are longer-term adverse impacts on earnings from unemployment experienced early in the employment lifecycle. Therefore, they clearly refute the view that the unemployed youth get trapped into unsteady and lowing paying jobs due to unemployment spells. However, it is evident in the theoretical model and empirical evidence, that these youths do not fully recover from the adverse effects of unemployment. It is recommended that the adverse effects of youth unemployment can be reduced through wage subsidy programs or subminimum wages and effective policies that encourage enhanced job training. However, the paper fails to test the effectiveness of its recommendations.

Marksoo (2011) investigates the impact of region of the residence and individual characteristics on the probability of being long-term unemployed in Estonia for the period 2000 – 2010. Changes in the structure of long-term unemployment and trends are examined

both during the economic recession and growth periods. The results of this study show that there was a significant variation in the incidence and duration of unemployment in the data. According to this study, changes in the demand structure of labour leading to structural unemployment was the major reason for long term unemployment in Estonia. The findings show that eastern Estonian residents, ethnic minorities, older people and people with low educational background have a higher probability of being long-term unemployed. This validates the author's hypothesis. The author claims that both supply-side and demand-side measures should be implemented in reducing the duration of unemployment.

Leuvensteijn & Koning (2000) evaluate the accuracy of profiling for UI and SA beneficiaries by estimating the duration models. The authors attempt to separate and assess the importance of the individual effects and 'indirect' sorting effects on the duration dependence effect. Their findings show that 'sorting effects' explained by the observed individual characteristics affect the job finding rate only in a limited way while the duration dependence effects at the individual level were found to be more important. The authors could delve further by analysing the motivation of the unemployed. This may improve the accuracy of profiling techniques, reducing deadweight risks.

The authors also point out that targeting specific groups alone while profiling bears a great risk of long-term unemployment for those unemployed that are (initially) classified as having good job prospects, as they might have received less training. Since the job finding rate deteriorates significantly due to 'individual duration effects', the authors suggest that labor market policies should not only rely on profiling at the start of an unemployment spell, but also on supplemental policies, such as encouraging search activities of all workers that have spent a certain length of time in unemployment.

Connell et al (2012) develops a statistical profiling model of long-term unemployment risk in Ireland. It is evident in the findings that factors such as a recent history of long-term unemployment, age, number of children, relatively low levels of education, literacy/numeracy problems, location in urban areas, lack of personal transport, low rates of recent labour market engagement, spousal earnings and geographic location all significantly affect the likelihood of remaining unemployed for twelve months or more. The results show

that community-based employment schemes for combating long-term unemployment were inefficient as participants re- entering the register typically experienced extended durations of unemployment (Cornell et al, 2012).

This paper points out that the accuracy of the profiling model depends on the labour market interventions associated with it. This study concludes that compared to the current nondiscriminatory intervention approach prevalent in Ireland, statistical profiling can offer a lot more in terms of efficiency and equity. However, the authors also acknowledge that the development and delivery of effective active labor market programs is essential to fully exploit the potential of profiling.

Weber (2011) covers the area of the role of profiling systems for effective market integration. The author claims that profiling, i.e. assessment performed by PES staff and through the use of IT and statistical tools for profiling can play a vital role in the personalisation of PES services. At national level and in many PES, various approaches have been developed to collect these "soft skills" through IT based profiling tools. According to this study, the rise in the number of diverse and complex statistical profiling tools which were developed to enable an early diagnosis of risk of long-term unemployment and customer segmentation, can be linked directly to decision making involving resource allocation such as the frequency and intensity of personal interviewing among other measures. Statistical profiling tried to identify job seekers' "needs" in terms of risk (e.g. risk of remaining/becoming long-term unemployed), which is related to client characteristics (e.g. gender, age, occupation, work experience etc.) (OECD, 2018).

Weber (2011) also argues that a good statistical profiling tool should not only take into account hard factors, but also include "soft" factors such as motivation, health and so on, as well as demand-side data (regional labour market information). She also emphasizes that the statistical profiling tools need to be simplistic and accurate as too much complexity of the tools would result in increase in the administrative burden, substantial documentation and workload for PES staff. Other conclusions include the effectiveness of the profiling system for resource allocation depends on the flexibility of the rules and decisions about allocation of Active Labor Market Policies (ALMPs). The reliability of methodology and

underlying data is critical for reliable output, as evidenced by the reliability of the profiling tools in Finland and Netherlands.

Loxha and Morgandi (2014) develop a new analytical framework, providing conceptual insights about typological features that differentiate profiling systems. This study analyzes and compares job seeker profiling methods adopted by the public employment services (PESs) of Organization of Economic Co- operation and Development (OECD) member countries. According to this study, profiling should enable PESs to segment jobseekers into groups with similar risk of work-resumption, and in turn to determine their level of access to different levels of treatment. It shows that PESs rely to a varying extent on (i) case worker discretion and on (ii) data intensive approaches. This study further points out that if PESs allocate interventions on a first-come-first-serve basis according to broad eligibility criteria (age, unemployment duration), it would either induce deadweight loss or result in delayed treatment. The outcomes for allocation from case manager's judgement depend strongly on the available time and capacity of case managers. This study provides an alternative approach that involves exploiting data about jobseekers to determine the probability of work-resumption according to a statistical model, which then allows the identification of customers that are most likely in need of active labor market interventions. According to this study, statistical profiling should be a suitable tool to maximize the impact of their scarce resources for PES in emerging economies which suffer from limited case management experience and high customer load.

Koen et al. (2013) examines if employability contributes to finding reemployment among the long term unemployed. The authors further investigate whether reemployment interventions facilitate the development of employability among the long term unemployed. The findings of this study show that employability can foster job search and the chance on finding reemployment for the unemployed despite the barriers of long-term unemployment. Based on their findings, the authors suggest that reemployment services should aim at assessing and fostering people's employability first, in order to assist the long-term unemployed in searching for and finding reemployment. One of the drawbacks of this study is that the authors used a broad categorization of reemployment interventions versus no

intervention, which put constraints on investigating links between receiving an intervention and the development of employability. Despite this drawback, their findings confirm the need for reemployment policies that simultaneously incorporate overcoming barriers and promoting employability among the long term unemployed. The authors argue that more person-centered interventions are needed to improve the effectiveness of reemployment policies to assist the long term unemployed.

The profiling systems of five European countries and the pros and cons of each system, by Barnes, Wright (2015), are presented in the table below:

Table 1: Profiling systems of five European countries

Country and Type of Profiling	Advantages	Disadvantages
France Type of Profiling: Caseworker based Profiling	● French caseworker-based profiling system allows for specialisation among caseworkers. ● This type of profiling system allows caseworkers to work exclusively with jobseekers of a specific profile-type, thereby allowing them to specialise and develop a strong expertise in supporting the needs of a specific category of jobseeker.	● Blanket approach to early intervention, wherein, jobseekers are obliged to wait three months following their initial registration and diagnosis interview before meeting with their personal caseworker. ● Although this period might be adequate for most jobseekers, those needing help with their job-search and facing higher risk of long-term unemployment

		would benefit from receiving support at an earlier stage.
Germany Type of Profiling: Soft-profiling	● VerBIS allows the caseworker to link information on regional labour market opportunities to the jobseeker's profile based on their competencies. ● The German profiling system is designed to capture generic and soft skills in the assessment process which is favourable for jobseekers who do not have formal qualifications. ● The VLM platform (VerBIS, Jobboerse and JobRobot) facilitates a two-way 'matching' of jobseekers to job vacancies. ● Potential employers do not know about the specific profile (Profillage) of a jobseeker that has been matched to their job vacancy.	● The customised or personalised services by the German PES is an expensive method of service delivery. ● Since the final decision on the services provided to jobseekers is made by caseworkers and profiling is largely dependent on the 'human element' of subjective assessment, caseworkers need to be highly skilled and that they will require a high level of training and support. ● Potential employers may be prejudicial to the jobseeker if the jobseeker has been referred to them by PES under UB II,since they are now aware that the jobseeker has been unemployed for more than 12 months.
Ireland Type of Profiling:	● Cost effective use of resources by calibrating the	● Risk of jobseekers providing false information due

Combination of statistical profiling and caseworker discretion	intensity of the support based on a jobseeker's risk of becoming long-term unemployed. ● Jobseekers at high and medium level risk will receive faster and more intense support than those with a low risk of becoming long-term unemployed.	to fear of affecting their chances of receiving future benefit payments. ● Jobseekers may provide a false subjective assessment of their health to influence the possibility of claiming for a disability allowance in the future. ● Ireland's statistical profiling tool does not support sustainable labour market attachment which affects seasonal workers as they are categorically placed as low risk group of becoming long term unemployed. ● PEX (Probability of Exit tool),the statistical profiling tool, only determines the engagement path of the jobseeker but fails to orient jobseekers towards activation services. Such a decision is at the complete discretion of the caseworker.
Netherlands Type of Profiling:	● The profiling instrument will help the Dutch Institute for Employee Benefit Schemes or	● Jobseekers should possess basic level of IT-literacy to make use of the Work Profiler. This

Statistical Profiling	UWV shift towards digital service delivery, while achieving the current targets set by the government in terms of budget reduction. ● The Work Profiler allows the caseworker to get information about the jobseeker before meeting the jobseeker which allows the caseworker to adapt and tailor support to the specific needs identified by the profiling instrument. ● Jobseekers are assessed based on wide range of factors, including soft factors. This prevents them from being wrongfully assigned to a certain group based on few stereotypical' variables.	could be a barrier for a segment of the client group which will not be able to participate in the new system (about 10-20 per cent of all jobseekers). ● Since caseworkers do not lead jobseekers re-integration process into the job market, the jobseeker is given increased ownership of their job-search which could act as a deterrent to some jobseekers in the process of finding employment without the support of caseworkers.
United Kingdom (UK) Type of Profiling: Soft-profiling	● Caseworkers play a crucial role in identifying those jobseekers who are in need of more intensive support and/or referral to specific interventions.	● The allocation of support by caseworkers could be inequitable since the process relies on the caseworker's 'profile' decisions to identify the optimal support that can be provided to the jobseeker

	● This approach is relatively cost effective and therefore, very advantageous	● This process also relies on the jobseeker providing accurate information upon which a 'profile' can be made, since providing incorrect information would affect the services offered.

Matty (2013) sheds further light on the profiling systems of UK by following the development of a prediction model to measure the likelihood of a new Jobseeker's Allowance (JSA) claimant reaching long term unemployment (LTU). The author highlights three limitations of the existing model. The first limitation relates to cost-benefit analysis, where the incorrect prediction of the model could lead to false positives. This could prove costly as candidates would be sent for unnecessary expensive interventions. An operational test of a segmentation approach is proposed to manage this limitation. The author regards the input data dependency on a specific one-off survey to be the second limitation, as it results in operational issues. To resolve this, the assessment of an individual needs to be done through a 'real time' IT solution or through a process solution that required new claimants to provide relevant information at the point of first contact, before any initial interview with an Adviser. It would also be necessary to test whether the change in setting and mode of data collection affected model accuracy. The last limitation highlights the need for a dynamic model to account for the variability in input data in terms of young people and older people datasets, timing of data collection, where it happens at any point in the claim process. Therefore, the author proposes to keep the model under continuous review.

Apart from studying the profiling systems in Europe, the profiling systems in other continents also provide special insights. Lipp (2005) tracks the evolution of the job profiling system in Australia and delves into the effectiveness of the country's current profiling system, the Job Seeker Classification Instrument (JSCI). The author tests the predictive

power of the JSCI by analysing the statistical output of the Logistic Regression model. The main consideration is that Logistic Regressions performed on cross-sectional data have low predictive power. Overall, the JSCI is known to deliver a good job in achieving its objective, on the back of highly accurate input data fed into the model. The author believes that the primary driver behind the success of the JSCI to date has been to stream job seekers to appropriate forms of assistance and levels of funding depending on the level of disadvantage. This results in not only reduced deadweight costs but also provides assistance to 'at risk' job seekers early in their unemployment spell before their barriers to employment become entrenched. The author concludes that Australia is effectively providing employment assistance through a focus on early intervention strategy, leading to the development of the JSCI. However, it is noted that Australia has to also use this performance information to provide case managers actionable insights on how best to effectively intervene and support individual job seekers in a cost-effective manner.

To conclude the literature review, it is evident that at a macro level, every country adopts a unique style of profiling system based on certain demographic characteristics and drivers of biases in decision-making in that country. The research papers reviewed in this review have thoroughly covered the main uses of a profiling system to tackle unemployment. Overall, the authors of the selected papers have provided several approaches to profiling in different countries. These studies have also provided an analytical framework to understand the trade-offs involved in the adoption of different profiling typologies. The in-depth illustration of statistical profiling methods and their corresponding selective case studies of the application of profiling tools to address policy priorities can be instrumental in tackling the issue of long- term unemployment through profiling. Based on the conclusions provided on the profiling systems of various countries and on further studying the specific characteristics, key decision-making drivers and demographic landscape of Estonia, the existing literature lays a solid foundation for the development of profiling system in Estonia.

4. METHODOLOGY

This research aims to test if modern statistical ensemble methods such as Random Forest or Gradient Boosting Machine can outperform simple Logistic Regression Models to predict long term unemployment.

4.1 Logistic Regression Model

The first classification algorithm in this research is Logistic Regression, which is a special case of linear regression where the outcome variable is categorical.

Logistic Regression can be modelled as:

$$\log(p/(1-p)) = b0 + b1\ x1 + b2\ x2 + \ldots bn\ xn$$

In the above equation, p is the probability of being registered unemployed for at least 360 days and x1, x2 … xn are the predictor variables which include variables like Age, Gender, County of Residence, Duration of last employment, etc. Logistic Regression model then generates the coefficients for these predictors, where higher coefficient indicates a higher chance of being registered unemployed for at least 360 days. Z is the total sum of the individual coefficients, where $Z= b0 + b1\ x1 + b2\ x2 + b3\ x3 \ldots bn\ xn$.

The final equation for the Logistic Regression model is given by:

$$p = \exp(Z)/(1+\exp(Z))$$

Here, p is the probability of being registered unemployed for at least 360 days and Z represents the overall score.

After splitting the dataset into Training (80 %) and Validation (20%) datasets, Logistic Regression is estimated on the Training dataset to predict the probability of being long term unemployed .The binary outcome variable is the dummy variable 'registered unemployed for at least 360 days' indicated by 1 when the person has been registered unemployed at least 360 days and 0 otherwise.

4.2. Random Forest Model

The second stage of this thesis involves estimating the Random Forest model. Random Forest is a supervised learning algorithm that builds multiple decision trees using 'bootstrapping' or 'bagging' method and provides the mode output of individual decision trees for prediction. Each individual decision tree is constructed from a random subset of the training data. Unlike single decision trees which suffer from high variance and high bias, Random Forest is a bagging ensemble model that fits the multitude of decision trees on various sub-samples of the dataset through bootstrap aggregation to control over-fitting and improve the predictive accuracy of the model.

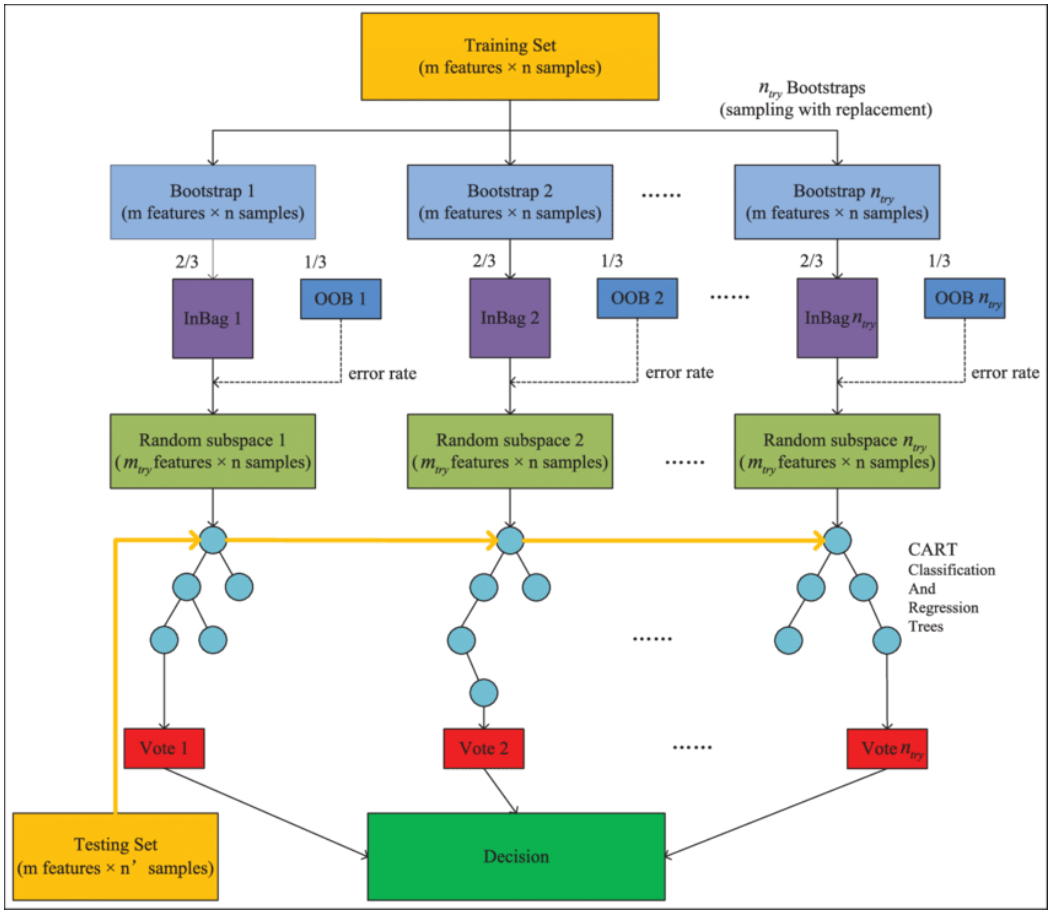

Figure 1: Workflow of the Random Forest Algorithm (Han et al, 2017)

Breiman's Random Forest estimation involved tuning of two key parameters, i.e. number of variables randomly sampled at each stage (mtry) and number of trees (ntree). At

each split within the tree, the Random Forest model uses a subset of variables to reduce the model's bias towards highly influential variables. The prediction error of Random Forest is measured through the Out-of-bag (OOB) estimate which uses bagging or bootstrap aggregating to sub-sample the training dataset. Firstly, I create a Random Forest model with default parameters. Then, I fine tune the Random Forest model by changing the number of trees (ntree) and the parameter (mtry), which is, the number of variables randomly sampled at each stage to minimize the OOB error rate.

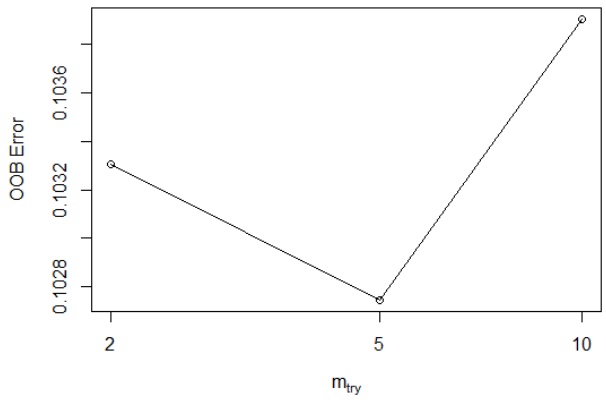

Figure 2: RF Out of Bag Error graph

As seen in the figure above, the OOB (Out Of Bag) Error was the lowest for 'mtry'= 5. Therefore, 400 trees(ntry) and 5 variables to be sampled at each stage(mtry) was chosen to be the optimal combination of hyperparameters in estimating Random Forest.

4.3 Gradient Boosting Machine Model

The third stage of this thesis involves estimating Gradient Boosting Machine (GBM), which is a boosting ensemble machine learning method used to solve regression and classification problems. Gradient Boosting Machine model is estimated in this thesis through the R package 'gbm', which is an implementation of the extensions of Freund and Schapire's AdaBoost algorithm and Friedman's Gradient Boosting machine. GBM's learning procedure involves a series of series of combinations of additive models (weak learning), estimated

iteratively resulting in a stronger learning model (stronger learning). GBM constructs an ensemble of shallow and weak successive trees with each decision tree learning and improving on the previous, unlike the Random Forest model estimated earlier, which involved construction of deep independent decision trees. Since GBM is sensitive to overfitting, it involves tuning of several hyperparameters to avoid overfitting of the Training data. Some of the key hyperparameters include number of trees, depth of trees and learning rate or shrinkage. On tuning the hyperparameters while estimating GBM on the Training data and cross validation on Testing data, Number of trees: 840, Depth of trees: 5 and shrinkage rate :0.1 was the optimal combination of the hyperparameters.

5. DATA

The data was extracted from the EUIF (Estonian Unemployment Insurance Fund) database which includes 256703 unemployment spells from 2014 to 2017. The data comprises of individual level anonymised historical data with information on socioeconomic characteristics (age, gender, education, place of living, etc.), labour market history, labour market services received, information on the duration of unemployment. Out of 256,703 unemployment spells, 26943 or 10% of the individuals were unemployed for at least 360 days or long term unemployed. On the other hand, 229760 or about 90% of the overall unemployed individuals were not long term unemployed. 51% of the registered unemployed individuals were male while the remaining 49% were unemployed females. The highest proportion of registered unemployed individuals belonged to the age group [25-30] and [20-25] while only 2%of registered unemployed belonged to the age group [60-65]. One of the key variables on labour market history is 'Duration of last employment'. The highest proportion of unemployed persons (31%) belonged to the class with the shortest duration of last employment (0-89 days).On the other hand, the proportion of registered unemployed persons decreases with higher durations of last employment, with just 5% of them belonging to the class of 3600 days & above in terms of last employment duration. Under field of education, 14% of the registered unemployed belonged to educational background 'service'

and 19% belonged to educational background in ‘manufacturing and construction’ while ‘humanitarian’ accounted for 2% of the registered unemployed. One of the drawbacks of this data is the absence of variables on soft skills such as motivation, problem solving skills, self-confidence, etc. which could be vital in predicting long term unemployment through statistical profiling of the unemployed (Weber, 2011).

6. RESULTS AND DISCUSSION

6.1 Accuracy Comparison of models

6.1.1 Accuracy comparison through Confusion Matrix

To evaluate and compare the prediction accuracy of the three models in this thesis, namely, the Logistic Regression model, Random Forest model and Gradient Boosting machine model, the Confusion matrix is constructed for each model on the testing dataset. Confusion matrix helps to visualise the performance of these classification models by comparing the actual outcomes to the predicted outcomes. Before constructing the confusion matrix, I adjust the classification threshold which has a default value of 0.5. This threshold is used to convert the predicted probabilities into class predictions, wherein an event is predicted to happen if the probability that the event occurs is greater than the threshold value and otherwise if the probability is lower than the threshold value. In the confusion matrix below, 0 indicates that the person is not long term unemployed while 1 indicates that the person is long term unemployed.

Although adjusting the threshold value affects the overall accuracy of the model’s predictions, it helps to determine whether to maximize or minimise the Sensitivity or Specificity for better prediction. Accuracy, Sensitivity, Specificity, Precision, False Positive Rate and Negative Predictive Value are the metrics used for in depth comparison of the prediction accuracy of these models. Accuracy shows the overall accuracy of the model’s

prediction and is calculated by the formula, Accuracy = (True Positives + True Negatives)/(True Positives + True Negatives + False Positives + False Negatives). Sensitivity or True Positive rate indicates how many positive values, out of all the positive values were correctly predicted and is calculated by the formula, Sensitivity = (True Positive/True Positive + False Negative).Specificity or True Negative Rate indicates the actual negatives that are correctly predicted by the model and is given by the formula, Specificity= (True Negative /(True Negative +False Positive).The formula for False Positive Rate is (1 - Specificity). In order to interpret the results from confusion matrix, the table below explains the basic terminology involved in the prediction analysis of the models.

Table 2: Explanation of prediction metrics

True Positive	The number of persons who are actually long term unemployed and correctly predicted as long term unemployed
True Negative	The number of persons who are not actually long term unemployed and correctly predicted as not long term unemployed
False Positive	The number of persons who are actually long term unemployed and incorrectly predicted as not long term unemployed
False Negative	The number of persons who are not actually long term unemployed and incorrectly predicted as long term unemployed

Table 3: Logistic Regression Model Confusion Matrix

Threshold: 0.6		
Actual value	(0) Predicted value	(1) Predicted value
0	45831 (TP)	47 (FN)
1	5349 (FP)	114 (TN)

Logistic Regression Model Threshold :0.3		
Actual value	(0) Predicted value	(1) Predicted value
0	43965 (TP)	1913(FN)
1	4168 (FP)	1295(TN)

Table 4: Logistic Regression Model Performance Metrics

Logistic Regression Model	Threshold : 0.3	Threshold : 0.6
Accuracy	0.88	0.894
Precision (TP/(TP+FP))	0.913	0.8954
Sensitivity or True Positive Rate(TP/(TP + FN))	0.958	0.998
Specificity (TN/(TN + FP))	0.237	0.02
False Positive Rate (FP/(FP+TN))	0.76	0.987
Negative Predictive value	0.40	0.70

The confusion matrix for the Logistic Regression Model with threshold value at 0.6 shows us that the number of persons who were actually long term unemployed and were correctly predicted to be long term unemployed or True Negatives, in this case, by the Logistic Regression model is 114 persons. Also, the number of persons who were not long term unemployed and were correctly identified as not long term unemployed or True Positives is 45831.The overall accuracy of the model is 0.89 or 89 percent accurate. The True Positive Rate or Sensitivity of the model is almost 100 percent. Due to the inverse relationship between Sensitivity and Specificity, the Logistic Regression model's confusion matrix shows very low Specificity of 2 percent. Also, the False Positive rate, which indicates how many negative values, out of all the negative values were incorrectly predicted was very high (99 percent). The number of persons who were actually long term unemployed but were incorrectly predicted to not being long term unemployed or False positives is 5349.

Since the main aim of this thesis is profiling or finding those unemployed who have a higher risk of being long term unemployed, it is imperative to tune the model's prediction by setting a threshold value that detects correctly the number of unemployed persons who are at risk of being long term unemployed. Therefore, it is critical to reduce the number of False positives which can be achieved by adjusting the threshold value of the model's predictions. In order to capture more unemployed persons with the risk of being long term unemployed, the Specificity of the model needs to be improved. This is achieved by reducing the threshold value of predicting long term unemployment from 0.6 to 0.3.

The results above show a negligible decrease in the overall accuracy of the Logistic Regression model's predictions of 1 percent after reducing the threshold from 0.6 to 0.3. The Precision of the model increases slightly when the threshold value is 0.3. The Sensitivity of the model decreases from 99 percent to 95 percent at threshold value 0.3. This decrease in Sensitivity could be attributed to the fall in True positives or the number of persons who were not long term unemployed and were correctly identified as not long term unemployed from 45831 to 43965. However, the main priority is to reduce the number of long term unemployed persons who were incorrectly predicted as not long term unemployed and this is accomplished by reducing the threshold value from 0.6 to 0.3 .This is evident in the Logistic

Regression model's Confusion matrix under threshold 0.3 where the False positive is 4168 persons compared to 5349 persons under the threshold value 0.6.This reduction in False Positives increases the Specificity of the model considerably from 2 percent to 23 percent, thus, improving the model's ability to capture more persons who have a risk of being long term unemployed that were initially incorrectly identified as not being long term unemployed. On the other hand, the Negative Predictive value falls from 70 percent to 40 percent after the change in the threshold value as it is inversely related to Precision.

Table 5: Random Forest Confusion Matrix

Threshold: 0.6		
Actual value	(0) Predicted value	(1) Predicted value
0	45758	120
1	5214 (FP)	249 (TN)
Threshold: 0.3		
Actual value	(0) Predicted value	(1) Predicted value
0	42780 (TP)	3098 (FN)
1	3640 (FP)	1823 (TN)

Table 6: Random Forest Performance Metrics

Random Forest Model	0.6	0.3
Accuracy	0.896	0.868
Precision (TP/(TP+FP))	0.897	0.921
Sensitivity or True Positive Rate (TP/(TP + FN))	0.997	0.93
Specificity (TN/(TN + FP))	0.04	0.33
False Positive Rate (FP/(FP+TN))	0.954	0.67
Negative Predictive value	0.674	0.370

The Random Forest confusion matrix under threshold value 0.6 shows overall Accuracy of 90 percent which is higher than the accuracy of Logistic Regression model under the same threshold value. While comparing the confusion matrix under threshold 0.6 of Random Forest and Logistic Regression model, the number of persons who were not long term unemployed and correctly predicted by Random Forest model as not long term unemployed is significantly greater than the True Positives of the Logistic Regression model. Similarly, the number of True Negatives is also greater in the Random Forest model's predictions, where Random Forest accurately predicted 249 long term unemployed persons while Logistic Regression model rightly identified 114 persons who were actually long term unemployed. Random Forest under threshold 0.3 also performed better in terms of Precision

compared to Logistic Regression model. Sensitivity of Random Forest model under threshold 0.6 is almost 100 percent while Specificity is just 4 percent. After changing the threshold value to 0.3, the confusion matrix shows a decrease in Accuracy by 2 percent, while showing an increase in the model's Precision. The Sensitivity of the Random Forest model decreases by 5 percent under threshold 0.3. The improvement in the Random Forest model's Specificity from 4 percent to 33 percent under threshold value 0.3 is better compared to the increase in Specificity under Logistic Regression model. The number of persons who were incorrectly predicted to not being long term unemployed while actually being long term unemployed (False Positives) reduces substantially from 5214 persons to 3640 persons. The number of False Positives is lowest under Random Forest model's confusion matrix under threshold value 0.3 and, therefore, achieves the goal of capturing more long-term unemployed persons. Therefore, the False positive rate reduces to 67 percent (threshold 0.3) from 95 percent (threshold 0.6).

Table 7: Gradient Boosting Machine Confusion Matrix

Threshold: 0.6		
Actual value	(0) Predicted value	(1) Predicted value
0	45724 (TP)	154 (FN)
1	5151 (FP)	312 (TN)
Threshold: 0.3		
Actual value	(0) Predicted value	(1) Predicted value
0	43795(TP)	2083 (FN)
1	3950 (FP)	1513 (TN)

Table 8: Gradient Boosting Performance Metrics

GBM Model	0.6	0.3
Accuracy	0.896	0.88
Precision (TP/(TP+FP))	0.898	0.91
Sensitivity or True Positive Rate(TP/(TP + FN))	0.996	0.954
Specificity (TN/(TN + FP))	0.05	0.276
False Positive Rate (FP/(FP+TN))	0.94	0.72
Negative Predictive value	0.669	0.42

Under threshold 0.6, Gradient Boosting machine model performed with the same accuracy as the Random Forest model. The results also show minimal differences in the performance of GBM and RF model in terms of precision, sensitivity and specificity. The confusion matrix under threshold 0.6 also indicate that the total number of true negatives and True Positives predicted by the GBM is higher compared to both, Random Forest and Logistic Regression model. However, the number of persons who were actually long term unemployed but incorrectly predicted by GBM model as not long term unemployed (false positives), was higher than Random Forest model under threshold 0.3.

Table 9: Performance Comparison

Models	LR :0.6	LR:0.3	RF:0.6	RF:0.3	GBM:0.6	GBM:0.3
Accuracy	0.89	0.88	0.90	0.87	0.90	0.88
Precision (TP/(TP+FP))	0.895	0.913	0.897	0.921	0.898	0.91
Sensitivity or True Positive Rate(TP/(TP + FN))	0.998	0.96	0.997	0.93	0.996	0.954
Specificity (TN/(TN + FP))	0.02	0.24	0.04	0,33	0.05	0.28
False Positive Rate (FP/(FP+TN))	0.99	0.76	0.954	0.67	0.94	0.72
Negative Predictive value	0.70	0.40	0.67	0.37	0.67	0.42

The above comparison between Logistic Regression model, Random Forest model and Gradient Boosting model in terms of Accuracy, Precision, Sensitivity and False Positive rates show modest differences in the performance of these models. However, the change in the threshold value from 0.6 to 0.3 offers a deeper insight through the subsequent changes in performance metrics like Accuracy, Specificity, False Positive Rate, etc. Additionally, it shows us that the model's ability to correctly predict long term unemployed persons can be tested by adjusting the threshold value for prediction. On close observation, the table above shows that Random Forest model experienced the largest decrease in False Positive Rate compared to Gradient Boosting Machine and Logistic Regression model after the change in threshold value, even though Accuracy decreased by 3 percent. Also, the Specificity of

Random Forest model is highest (33%) which shows that Random Forest was more effective in correctly predicting long term unemployment than Gradient Boosting Machine and Logistic Regression model and was capable of predicting the lowest number of False Positives.

6.1.2 Accuracy comparison through ROC AUC

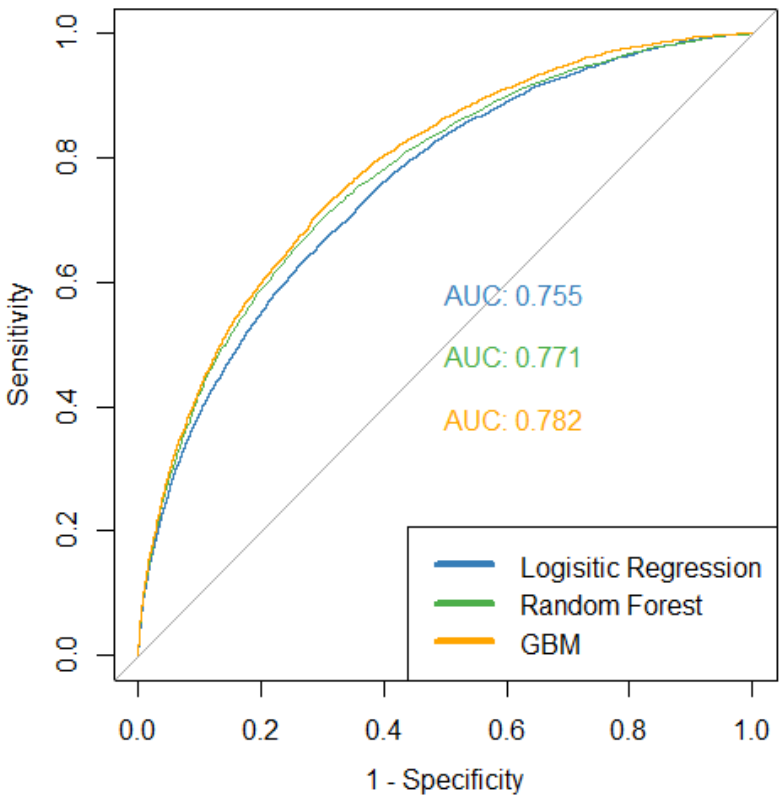

Figure 3: ROC AUC Graph

For further comparison of the performance of these models, the ROC graphs of the Logistic Regression model, Random Forest model and Gradient Boosting Machine are examined above. The ROC or Receiver Operating Characteristic is a graphical plot that is generated by plotting the Sensitivity or True Positive Rate against the False Positive Rate (1 - Specificity). The ROC curves above represent the proportion of correctly predicted or

identified long term unemployed persons against the incorrectly or misclassified proportion of long term unemployed persons. The Area under curve (AUC) is the percentage of area under the ROC curve and is a perfect performance metric for ROC curve that summarizes the performance of a classifier in a single number. GBM model had the highest AUC value of 0.78, while RF model performed slightly worse than GBM with an AUC score of 0.77. The Logistic Regression model scored the lowest in terms of AUC (0.75) compared to RF and GBM models.

6.2 Variable Importance

6.2.1 Variable Importance in Random Forest Model

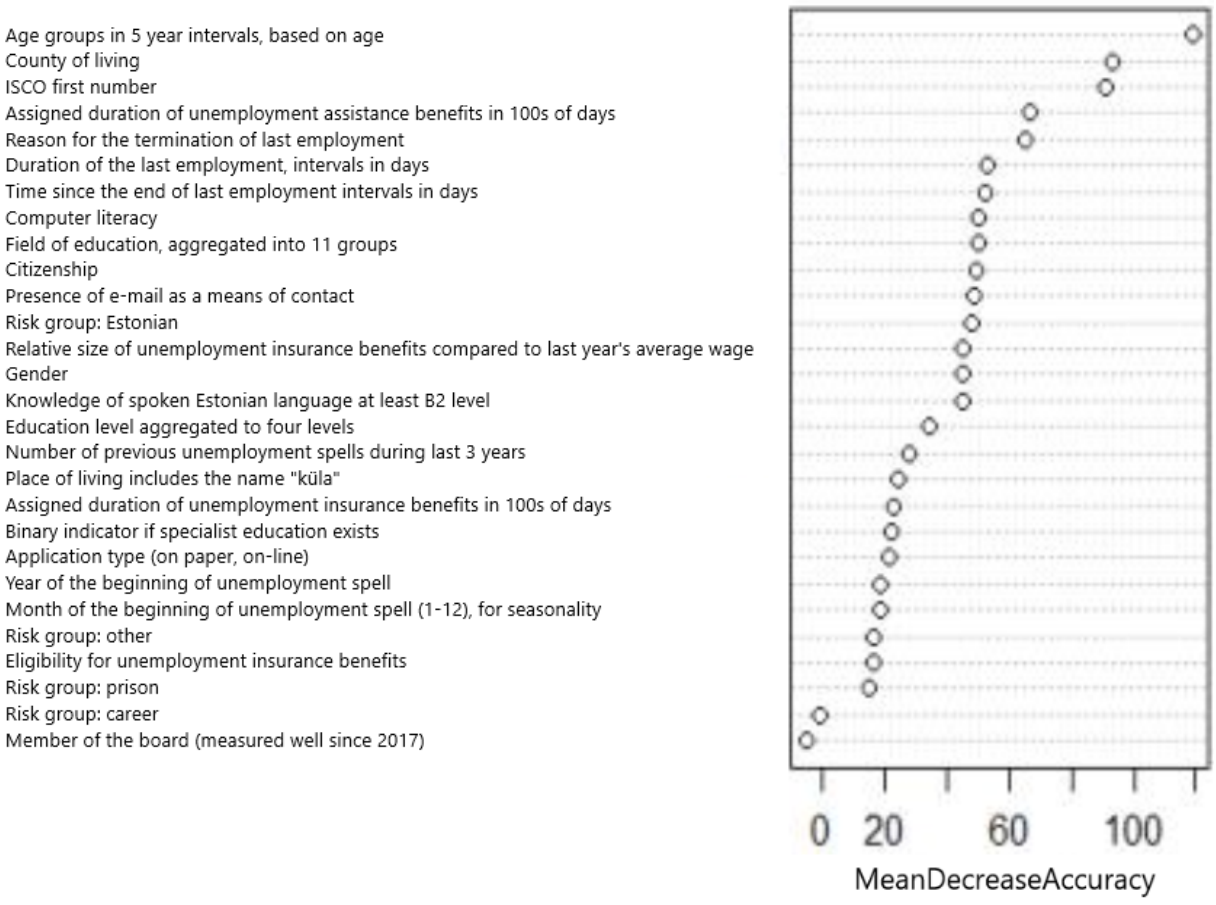

Figure 4: RF Variable Importance: Mean Decrease Accuracy

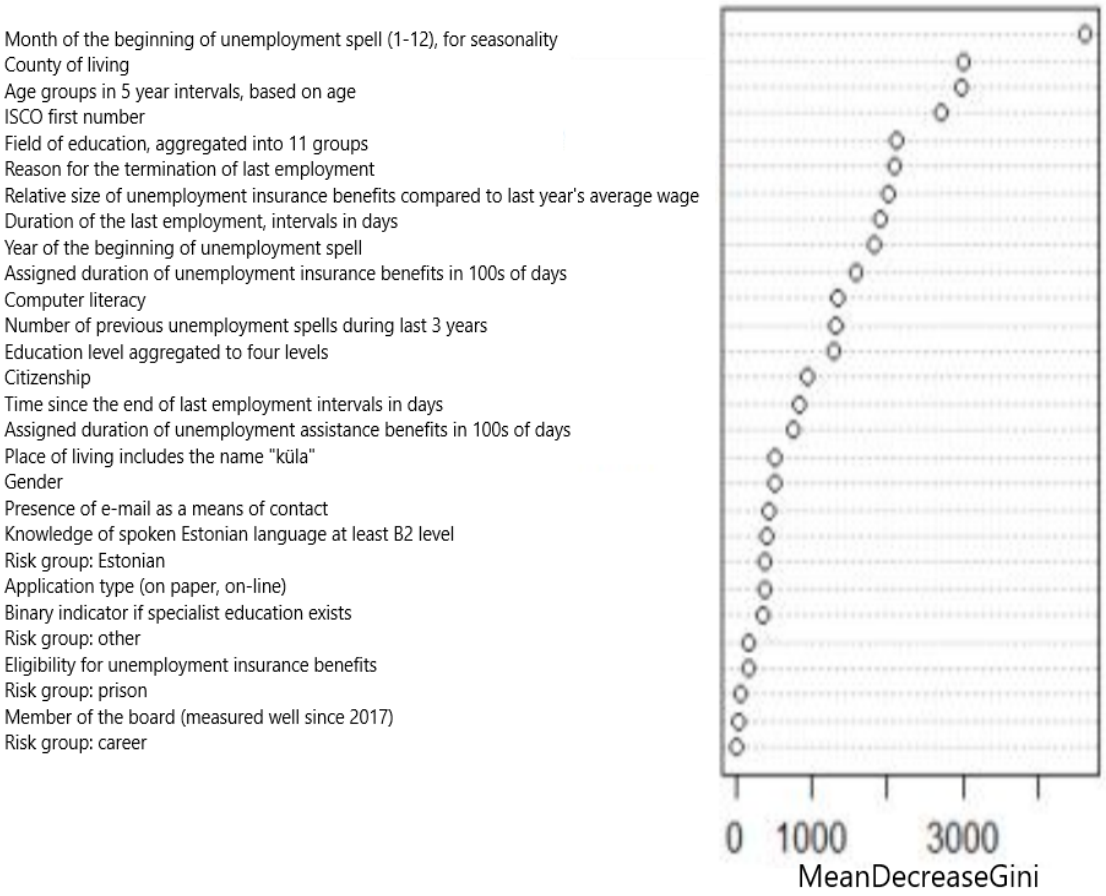

Figure 5: RF Variable Importance: Mean Decrease Gini

In order to check the predictive power of variables in the model, the Variable Importance graph above explains which variables significantly affect the outcome of the model. In the Random Forest model, there are two measures of importance given for each variable, namely, Accuracy Importance and Gini Importance.

Mean Decrease Accuracy shows how the Random Forest model's accuracy decreases when the variable is excluded from the model. The out of bag sample data which was not used during construction from each tree is used to calculate the specific importance of the variable. The prediction accuracy of the out of bag sample of the variable is measured before randomly shuffling the values of the variable. After shuffling, the decrease in prediction accuracy is measured to check the predictive power of the shuffled variable. Mean Decrease Gini is based on the decrease of Gini impurity when a variable is chosen to split a node. Since

splits are biased towards variables with several classes, the Gini importance could also be biased while calculating the importance measure of such variables.

'Age','County of living' and 'Reason for termination of last employment' are among the variables that show the highest predictive power according to the Mean Decrease Accuracy and Mean Decrease Gini graphs above.On the other hand, 'Member of the board', 'Eligibility of unemployment insurance benefits','Risk Group:Prison' and 'Risk Group: Carer' were the variables with lowest importance in the Random Forest model which is evident in the Mean Decrease Accuracy and Mean Decrease Gini graphs above.

Distribution of minimal depth and Mean

Now, through the 'randomForest Explainer' package in R, I calculate and plot the distribution of minimal depth of variables in the Random Forest model which offers more insight into the role of variables that impact the model's prediction. In Random Forest model, the minimal depth of a variable is a surrogate measure of predictiveness of that variable. The smaller the minimal depth, the more impact the variable has sorting observations, and therefore on the forest prediction.

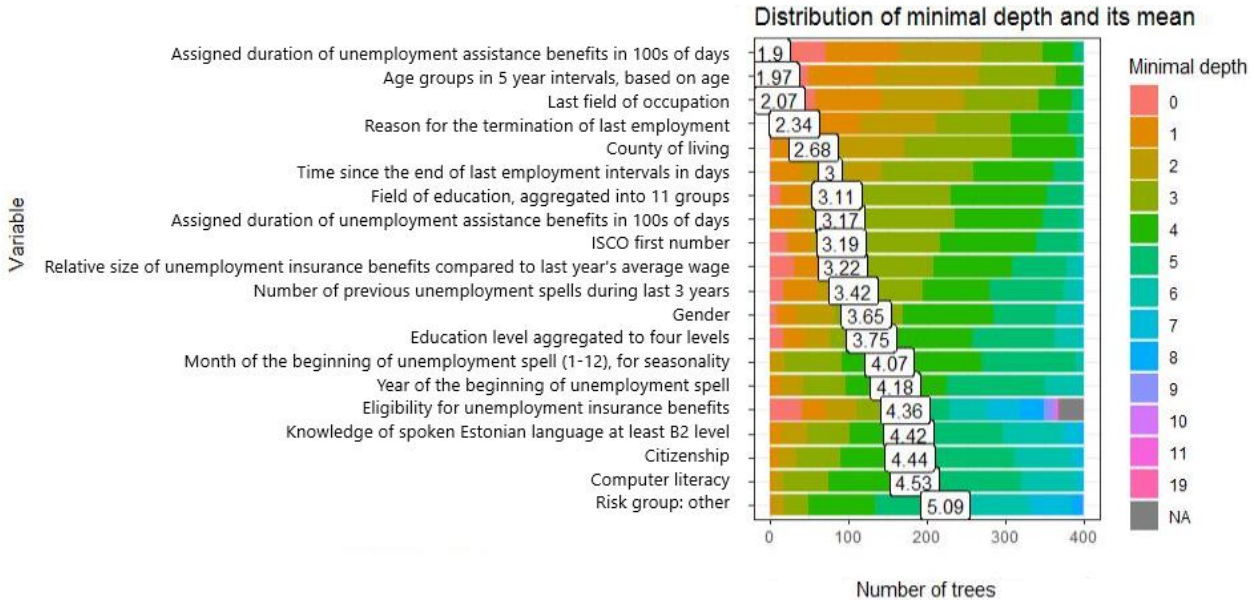

Figure 6: RF Distribution of Minimal Depth and Mean

The variables ‘Assigned duration of unemployment insurance benefits in 100s of days’, ‘Age groups’, ‘Duration of last employment’ were the variables with the lowest mean minimal depths and therefore could be considered as highly important variables.

Random Forest Multi-way importance plot

In the multi-way importance plot, x-axis represents ‘times_a_root’ which indicates the total number of trees in which the variable is used for splitting the root node while the y-axis represents the mean minimal depth. The inverse relationship between ‘time_a_root’ and ‘mean_minimal_depth’ is evident in the multi-way importance plot below. ‘Assigned duration of unemployment insurance benefits in 100s of days’,‘Duration of last employment’,’Reason for termination of last employment’, ‘Age groups’ were the variables with the lowest mean minimal depths and high ‘times_a_root’ and therefore were the most important variables.

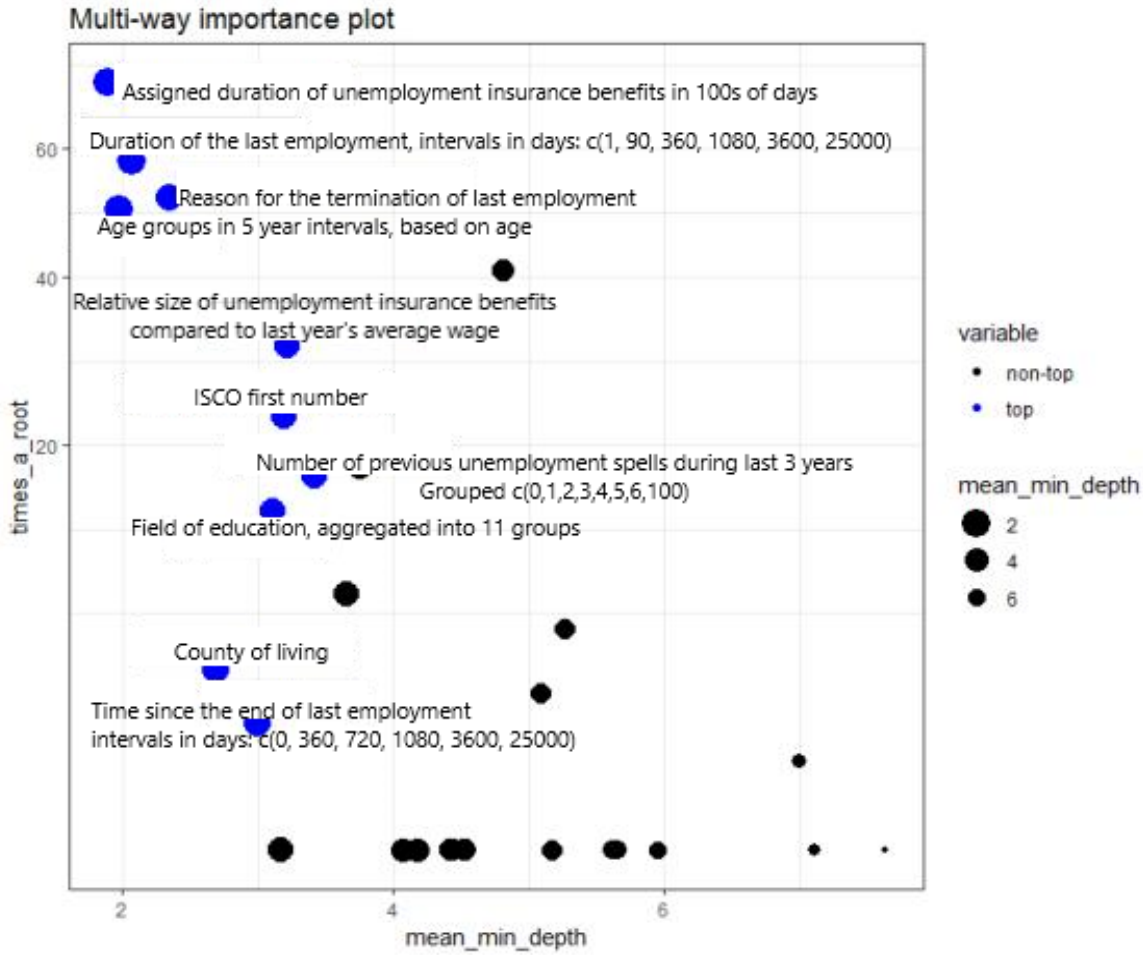

Figure 7: RF Multi-way Importance Plot

6.2.2 Variable Importance in Gradient Boosting Model

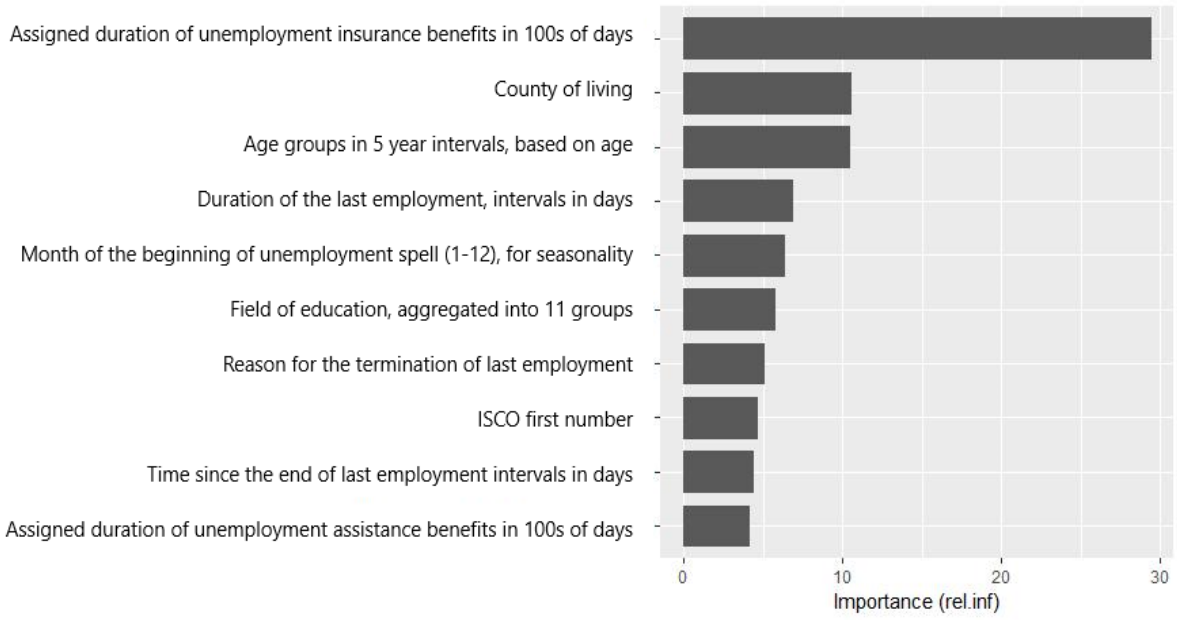

Figure 8: GBM Variable Importance

The method for calculating variable importance in GBM is similar to the Mean Decrease in Accuracy measure presented earlier in Random Forest model. Similar to the Distribution of Minimal Depth graph from Random Forest model, the variables 'Assigned duration of unemployment insurance benefits in 100s of days' and 'Age group, 'county of living' were the variables with the highest predictive influence on the GBM model.

According to the ROC-AUC graphs, Gradient Boosting machine performed better than Random Forest and Logistic Regression model. However, based on the prediction metrics from the confusion matrices, Random Forest was more effective in capturing the long term unemployed compared to Gradient Boosting machine and Logistic Regression model.

Although Random Forest and GBM models outperform Logistic Regression model in terms of accuracy prediction, the interpretability of Random Forest and GBM may not be reliable as the Variable Importance measures from these models could be biased towards

potential predictor variables that may vary in their scale of measurement or their number of categories.

7. CONCLUSION

This thesis investigates if modern statistical ensemble methods, such as Random Forest or Gradient Boosting machine can outperform simple Logistic Regression models, which has been used in profiling of long term unemployed earlier in Estonian studies (Trumm, 2018). The analysis in this thesis involved the estimation of Logistic Regression, Random Forest and Gradient Boosting machine models using the data extracted from the EUIF (Estonian Unemployment Insurance Fund) database consisting of 256,703 unemployment spells from 2014-2017. The prediction accuracy of these classification models is then evaluated under the threshold values of 0.3 and 0.6 on the basis of crucial performance metrics such as Accuracy, Precision, Sensitivity, Specificity and False Positive rate. ROC graphs were generated and the AUC of each model was calculated to further evaluate and compare the performance of the Logistic Regression, Gradient Boosting machine and Random Forest models. The Variable Importance graphs are also presented to interpret the level of influence of the variables on the Random Forest model and Gradient Boosting model's prediction. The variable importance graphs show that 'Assigned duration of unemployment insurance benefits in 100s of days', 'Age group', 'county of residence' were the variables with the highest predictive influence on the Random Forest and GBM model.

According to the ROC-AUC graphs, Gradient Boosting machine performed better than Random Forest and Logistic Regression model. The Specificity of Random Forest model is highest (33%) which shows that Random Forest was more effective in correctly predicting long term unemployment than Gradient Boosting machine and Logistic Regression model. The confusion matrix of Random Forest model also showed the lowest number of False Positives (incorrect prediction of long term unemployed). However, based

on the prediction metrics from the confusion matrices, Random Forest was more effective in capturing the long term unemployed compared to Gradient Boosting machine and Logistic Regression model. Although Random Forest and GBM models outperform Logistic Regression model in terms of accuracy prediction, the interpretability of Random Forest and GBM may not be reliable as the Variable Importance measures from these models could be biased towards potential predictor variables that may vary in their scale of measurement or their number of categories. Being the simplest machine learning algorithm compared to Random Forest and Gradient Boosting machines, Logistic Regression may not be very effective in its prediction accuracy when the data includes highly correlated variables or noisy data due to its lack of flexibility. On the other hand, estimation of Logistic Regression is easier and faster compared to complex models like GBM and Random Forest. Although Gradient Boosting model outperformed Logistic Regression and Random Forest in terms of accuracy prediction (ROC-AUC), training in GBM takes longer as trees are built sequentially. GBM estimation also involves tuning of several hyperparameters while Random Forest involves the tuning of the number of trees and number of features to be sampled. One of the crucial weaknesses of GBM and Logistic Regression models is that these models are more sensitive to overfitting in case of noisy data. Although Random Forests are less interpretable compared to Logistic Regression model and slightly lower than GBM in terms of prediction accuracy, Random Forests are not vulnerable to overfitting and therefore can be considered as suitable models for statistical profiling of long term unemployed.

The challenge is to overcome the inherent limitations facing the effectiveness of statistical profiling systems. Some of these limitations include data lags, a lack of accuracy, and a lack of transparency. As data represents the past, it could not be a reliable source to understand the present and predict the future. This leads to data lag issues. The second limitation related to data accuracy could be contributed by the data collection system constraints such as jobseekers incorrectly classified as high or low risk candidates. As all individuals are treated with the same set of parameters, the risk of statistical discrimination exists. The transparency issue arises from problems in examining or analysing the statistical model or algorithm used in the system. These endemic limitations can only be resolved by

striving to keep improving the design of these profile systems. Better data such as labour market history and soft skills and latest updates could help improve the reliability of the data output. Human error could creep in as caseworkers collect data on the jobseeker's risk of long-term unemployment. Therefore, cost-benefit analysis should be done to evaluate whether the additional costs of more caseworker resources would justify the benefits that the job seekers receive in terms of support. To conclude, it is recommended that countries use the output of their chosen statistical profiling model as one of the contributing factors and not the only indicator in their decision-making mechanism to tackle long term unemployment.

REFERENCES

Barnes, S-A., Wright, S., Irving, P., Deganis, I. (2015), "Identification of latest trends and current developments in methods to profile jobseekers in European public employment services". *Final report, Brussels: Directorate-General for Employment, Social Affairs and Inclusion*. European Commission.

Breiman, L. (2001), "Random Forests." Springer, pp. 5-32.

Friedman, J. (2001), "Greedy function approximation: A Gradient Boosting machine." *Ann. Statist.*, 29, 5, 1189-1232.

Han, T., Jiang, D., Zhao, Q., Wang, L., Yin, K. (2017), Comparison of Random Forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. *Transactions of the Institute of Measurement and Control*.

James, G., Witten, D., Hastie, T., Tibshirani, R (2013), "An Introduction to Statistical Learning." Springer, pp. 316–321.

Kelly, E., Mcguinness, S., J Oconnell, P (2012), "Transitions to Long-Term Unemployment Risk Among Young People: Evidence from Ireland". *Journal of Youth Studies*.

Koen, J., Klehe, U., van Vianen, A. (2013), "Employability among the long- term unemployed: A futile quest or worth the effort?". *Journal of Vocational Behavior*, 82, 37-48.

Leuvensteijn, M., Koning, P. (2000), "Duration Dependence in Unemployment Insurance and Social Assistance: Consequences of Profiling for the Unemployed." Research Memorandum, CPB Netherlands Bureau for Economic Policy Analysis, The Hague.

Lipp, R. (2005), "Job Seeker Profiling – The Australian Experience." Australian

Government – Department of Employment and Workplace Relations.

Loxha, A., Morgandi, M. (2014), "Profiling the Unemployed A Review of OECD Experiences and Implications for Emerging Economics". *Social Protection and labor discussion paper*, SP 1424. Washington, DC: World Bank Group.

Marksoo, Ü. (2011), "Long-term unemployment and its regional disparities in Estonia". Tartu Univerity Press, (Tartu: Tartu University Publishing House).

Matty, S. (2013), "Predicting likelihood of long-term unemployment: the development of a UK jobseekers' classification instrument." Working Paper No. 116. London, UK: Department for Work and Pensions.

Mroz, A., Savage, T. (2006), "The Long-Term Effects of Youth Unemployment." *The Journal of Human Resources*, 41, 2, 259-93.

OECD (2018), "Profiling tools for early identification of jobseekers who need extra support." Policy Brief on Activation Policies, OECD Publishing, Paris.

Trumm, E. (2018), "Profiling the Unemployed on the Registry Data of the Estonian Unemployment Insurance Fund." Faculty of Social Sciences, University of Tartu.

Weber, T., (2011), "Profiling systems for effective labour market integration". *Thematic Synthesis Paper*. European Commission. Directorate-General for Employment, Social Affairs and Inclusion

ANNEX

Data includes 256703 unemployment spells from 1.1.2014-17.11.2018

Date of extraction from EUIF database 17.11.2018. Date of final preparation 27.03.2019.

Variable	In Estonian	In English
pseudo_id	pseudo id korduvate isikute leidmiseks	pseudo id number
arvel_lopp_pohjus_toole	1-0 tunnus, kas lahkuti tööle	Reason for the end of unemployment spell related to re-employment or not (1-0)
arvelekuu	Arvele tuleku kalendrikuu	Month of the beginning of unemployment spell (1-12), for seasonality
aasta	Arvele tuleku aasta	Year of the beginning of unemployment spell
pikaajalineregtootu180	1-0	Registered unemployed at least 180 days, since the beginning of registration
pikaajalineregtootu360	1-0	Registered unemployed at least 360 days
hoivesseregtootusealguse st180	1-0	Has left unemployment to employment within 180 days
hoivesseregtootusealguse st360	1-0	Has left unemployment to employment within 360 days
sugu	Väärtused [1] "M" "N"	Gender
vanusgrupp	"(20,25]" "(15,20]" "(25,30]" "(30,35]" "(35,40]" "(40,45]" "(45,50]" "(50,55]" "(55,60]" "(60,65]" "(65,70]"	Age groups in 5 year intervals, based on age
haridustase4	"Tase1" "Tase2" "Tase3" "Tase4" "Tase5" "Teadmata"	Education level aggregated to four levels

erialane_haridus	"0" "1" "Teadmata"	Binary indicator if specialist education exists
maakond_act_ehak	Maakonnad	County of living
kyla	1-0	Place of living includes the name "küla"
risk_eesti_keel	1-0	Risk group: Estonian
risk_hooldaja	1-0	Risk group: carer
risk_vangla	1-0	Risk group: prison
risk_muud	1-0	Risk group: other
kodakondsus	"Eesti" "Määramata" "Muu" "Teadmata" "Venemaa"	Citizenship
et_keel_kones_minb2	1-0	Knowledge of spoken Estonian language at least B2 level
juhatuse_liige	1-0	Member of the board (measured well since 2017)
arvutioskus	"ALGTASE" "EKSPERDI_TASE" "KESKTASE" "SPETSIALISTI_TASE" "Teadmata"	Computer literarcy
avaldus_saabumise_viis	"ITP kaudu" "Kohapeal"	Application type (on paper, on-line)
emailolemas	1-0	Presence of e-mail as a means of contact
oppevaldkond	"ärindusõigus" "haridus" "humanitaaria" "IKT" "loodusteadused" "PM" "sotsiaalteadused" "teadmata" "teenindus" "tervisheaolu" "tootmineehitus"	Field of education, aggregated into 11 groups
viim_hoive_valdkond	[1] "ajakirjandus, toimetamine, tõlkimine" [2] "avalik haldus" [3] "dokumendihaldus, personalitöö, infotöö" [4] "ehitus" [5] "ehitusmaterjalide, keraamikatööstus"	Last field of occupation

	[6] "elektri- ja energiatootmine, elektrimehhaanika" [7] "elektroonika, automaatika" [8] "ettevõtte, organisatsiooni juhtimine, kvaliteedijuhtimine" [9] "film, teater, ringhääling" [10] "finants, raamatupidamine, statistika" [11] "foto- ja trükitööstus" [12] "haridus: huviharidus, muu haridustöö" [13] "haridus: juhid ja pedagoogid" [14] "iluteenindus" [15] "infotehnoloogia, telekommunikatsioon" [16] "jäätmetöötlus, prügivedu" [17] "kaevandamine, mäetööstus" [18] "kaitsevägi, päästeteenistus, korrakaitse" [19] "kaubandus, klienditeenindus" [20] "keskkond, maamõõtmine, maakorraldus" [21] "klaasitootmine, klaasitööstus" [22] "kunst, fotograafia, muusika" [23] "laevandus, lennundus" [24] "liigitamata lihttöö" [25] "logistika, varustamine, laondus (va sõiduki- ja tõstukijuhtimine)" [26] "loodus- või täppisteadus" [27] "metalli- ja masinatööstus"	

	[28] "metsandus, jahindus, kalandus" [29] "mööblitööstus" [30] "muu teenindus" [31] "muu tööstus, tootmine" [32] "muuseum, galerii, raamatukogu, arhiiv" [33] "õigus" [34] "plasti-, kummi-, ravimi jm keemiatoodete tööstus" [35] "põllumajandus, veterinaaria" [36] "puhastusteenindus, majapidamine jms" [37] "puidu- ja paberitööstus" [38] "sõiduki või masina mehhaanik, lukksepp vms" [39] "sõidukijuhtimine, tõstukijuhtimine" [40] "sotsiaal- või humanitaarteadus" [41] "sotsiaaltöö, hooldus, lastehoid" [42] "sport, vaba aeg, meelelahutus" [43] "taimekasvatus, loomakasvatus, aiandus" [44] "täppisriistade parandus, käsitöö" [45] "Teadmata" [46] "tekstiili-, naha- ja jalatsitööstus" [47] "tervishoid: arstid" [48] "tervishoid: farmatseudid, laboritöötajad, terapeudid" [49] "tervishoid: meditsiini õde" [50] "toiduainetetööstus" [51] "toitlustus, majutus, ürituste korraldus" [52] "turundus, avalikud suhted, müügikonsultatsioon"	

viim_hoive_isco_kood1k	"0" "1" "2" "3" "4" "5" "6" "7" "8" "9" "99"	ISCO first number
aasta_hoivest	"Kuni 1" "1-2" "2-3" "3-5" "5-" "Teadmata"	Time since the end of last employment intervals in days: c(0, 360, 720, 1080, 3600, 25000)
varasemadtootused	"[0,1)" "[1,2)" "[2,3)" "[3,4)" "[4,5)" "[5,6)" "[6,100]"	Number of previous unemployment spells during last 3 years Grouped c(0,1,2,3,4,5,6,100)
valjamaksedkuud24ante		Fraction of months with employment income during 24 months before becoming unemployed
viim_hoive_kestusgr	"0-89" "90-359" "360-1089" "1080-3599" "3600-" "Teadmata"	Duration of the last employment, intervals in days: c(1, 90, 360, 1080, 3600, 25000)
viim_hoive_lopp_pohjusgr2	"Katseaeg ebarahuldav" "KoondaminePankrotLikv" "Muu" "Poolte kokkuleppel" "Tähtajaline tööleping" "Teadmata" "Tööandjapoolne rikkumine" "Töötaja algatusel" "Töötaja süü"	Reason for the termination of last employment (nine categories)
maaratud_tt_kestus_100		Assigned duration of unemployment assistance benefits in 100s of days
first_daily_ratepos	1-0	Eligibility for unemployment insurance benefits
rel_first_daily_rate		Relative size of unemployment insurance benefits compared to last year's average wage
days_appointed_100		Assigned duration of unemployment insurance benefits in 100s of days

Logistic Regression Model Output

	Dependent variable:
	Long term unemployed
arvelekuu2	0.042
	(0.035)
arvelekuu3	0.091***
	(0.035)
arvelekuu4	0.144***
	(0.036)
arvelekuu5	0.189***
	(0.036)
arvelekuu6	0.009
	(0.036)
arvelekuu7	0.108***
	(0.036)
arvelekuu8	0.137***
	(0.036)
arvelekuu9	0.143***
	(0.034)
arvelekuu10	0.005
	(0.035)
arvelekuu11	0.058
	(0.036)
arvelekuu12	0.038
	(0.041)
aasta2015	0.022
	(0.023)
aasta2016	0.059***
	(0.023)
aasta2017	0.162***
	(0.023)
suguN	0.208***
	(0.019)
vanusgrupp(15,20]	-0.353***
	(0.063)
vanusgrupp(25,30]	0.398***

	(0.040)
vanusgrupp(30,35]	0.653***
	(0.040)
vanusgrupp(35,40]	0.812***
	(0.040)
vanusgrupp(40,45]	0.897***
	(0.041)
vanusgrupp(45,50]	0.976***
	(0.041)
vanusgrupp(50,55]	1.139***
	(0.041)
vanusgrupp(55,60]	1.257***
	(0.041)
vanusgrupp(60,65]	1.027***
	(0.057)
haridustase4Tase2	0.139**
	(0.061)
haridustase4Tase3	0.091***
	(0.027)
haridustase4Tase4	0.018
	(0.039)
haridustase4Tase5	0.134***
	(0.046)
haridustase4Teadmata	-0.490***
	(0.132)
maakond_act_ehak39	0.238**
	(0.103)
maakond_act_ehak45	0.598***
	(0.024)
maakond_act_ehak50	-0.307***
	(0.067)
maakond_act_ehak52	0.324***
	(0.050)
maakond_act_ehak56	0.339***
	(0.060)
maakond_act_ehak60	0.119***
	(0.042)
maakond_act_ehak64	0.610***
	(0.051)

maakond_act_ehak68	0.320***
	(0.034)
maakond_act_ehak71	0.312***
	(0.051)
maakond_act_ehak74	0.076
	(0.059)
maakond_act_ehak79	0.101***
	(0.030)
maakond_act_ehak81	0.819***
	(0.043)
maakond_act_ehak84	-0.212***
	(0.054)
maakond_act_ehak87	0.719***
	(0.043)
maakond_act_ehakTeadmata	-1.030
	(1.030)
kyla1	0.145***
	(0.021)
risk_hooldaja1	-0.147
	(0.201)
risk_vangla1	-0.769***
	(0.075)
risk_muud1	0.535***
	(0.049)
risk_eesti_keel1	-0.949***
	(0.047)
et_keel_kones_minb21	-1.030***
	(0.047)
kodakondsusMääramata	0.091***
	(0.028)
kodakondsusMuu	0.196***
	(0.055)
kodakondsusTeadmata	-0.899***
	(0.167)
kodakondsusVenemaa	0.105***
	(0.031)
juhatuse_liige1	0.188*
	(0.114)
arvutioskusEKSPERDI_TASE	0.127**

	(0.051)
arvutioskusKESKTASE	0.061***
	(0.020)
arvutioskusSPETSIALISTI_TASE	0.280***
	(0.041)
arvutioskusTeadmata	-0.196***
	(0.032)
avaldus_saabumise_viisKohapeal	0.042
	(0.027)
emailolemas1	-0.035
	(0.026)
erialane_haridus1	-0.861***
	(0.079)
erialane_haridusTeadmata	
oppevaldkondharidus	-0.295***
	(0.052)
oppevaldkondhumanitaaria	-0.156***
	(0.054)
oppevaldkondIKT	-0.095
	(0.063)
oppevaldkondloodusteadused	-0.117*
	(0.064)
oppevaldkondPM	-0.357***
	(0.048)
oppevaldkondsotsiaalteadused	-0.205***
	(0.067)
oppevaldkondteadmata	-1.180***
	(0.080)
oppevaldkondteenindus	-0.323***
	(0.034)
oppevaldkondtervisheaolu	-0.293***
	(0.065)
oppevaldkondtootmineehitus	-0.232***
	(0.033)
viim_hoive_isco_kood1k0	0.164
	(0.241)
viim_hoive_isco_kood1k1	0.338***
	(0.038)

viim_hoive_isco_kood1k2	0.277***
	(0.039)
viim_hoive_isco_kood1k3	0.353***
	(0.034)
viim_hoive_isco_kood1k4	0.337***
	(0.040)
viim_hoive_isco_kood1k5	0.125***
	(0.031)
viim_hoive_isco_kood1k6	0.103*
	(0.063)
viim_hoive_isco_kood1k8	0.006
	(0.031)
viim_hoive_isco_kood1k9	0.233***
	(0.027)
viim_hoive_isco_kood1k99	-0.560***
	(0.122)
aasta_hoivest1-2	0.455***
	(0.038)
aasta_hoivest2-3	0.613***
	(0.047)
aasta_hoivest3-5	0.792***
	(0.032)
aasta_hoivest5-	0.785***
	(0.050)
aasta_hoivestTeadmata	1.123***
	(0.129)
varasemadtootused[1,2)	0.076***
	(0.021)
varasemadtootused[2,3)	0.067**
	(0.029)
varasemadtootused[3,4)	-0.101**
	(0.040)
varasemadtootused[4,5)	-0.179***
	(0.059)
varasemadtootused[5,6)	-0.381***
	(0.093)
varasemadtootused[6,100]	-0.740***
	(0.122)
viim_hoive_kestusgr90-359	-0.121***

	(0.024)
viim_hoive_kestusgr360-1089	0.058**
	(0.028)
viim_hoive_kestusgr1080-3599	0.162***
	(0.030)
viim_hoive_kestusgr3600-	0.515***
	(0.035)
viim_hoive_kestusgrTeadmata	
viim_hoive_lopp_pohjusgr2KoondaminePankrotLikv	0.222***
	(0.047)
viim_hoive_lopp_pohjusgr2Muu	0.798***
	(0.059)
viim_hoive_lopp_pohjusgr2Poolte kokkuleppel	-0.125**
	(0.051)
viim_hoive_lopp_pohjusgr2Tähtajaline tööleping	-0.073*
	(0.044)
viim_hoive_lopp_pohjusgr2Teadmata	0.191***
	(0.049)
viim_hoive_lopp_pohjusgr2Tööandjapoolne rikkumine	0.438***
	(0.110)
viim_hoive_lopp_pohjusgr2Töötaja algatusel	-0.181***
	(0.051)
viim_hoive_lopp_pohjusgr2Töötaja süü	0.073
	(0.061)
first_daily_ratepos1	-1.094***
	(0.053)
rel_first_daily_rate	-0.005
	(0.039)
days_appointed_100	0.743***
	(0.017)
maaratud_tt_kestus_100	0.342***
	(0.009)
Constant	-2.481***
	(0.121)
Observations	205,362
Log Likelihood	-59,368.240
Akaike Inf. Crit.	118,956.500

==========
==========
Note: *p<0.1; **p<0.05; ***p<0.01

Non-exclusive licence to reproduce thesis and make thesis public

I, Gratian Allan

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

Profiling of Registered Unemployed

Supervised by Andres Võrk

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Gratian Allan
23/05/2019