Predicting Long-Term Unemployment:

A Comparison Between Logistic Regression and Kernel SVM

Seminar Thesis

'Master Seminar on Econometrics'

Prof. Dr. Martin Biewen

Marian Rümmele

January 22, 2021

Malik Stromberg
Fichtenweg 26 / 305
72076 Tübingen
5396072
Data Science in Business and Economics

Contents

1 Introduction **1**

2 Context **2**

3 Method **4**

3.1 Logistic Regression 6

3.2 Support Vector Machine 7

3.3 Model Evaluation 9

4 Data **10**

5 Results **13**

5.1 Feature Selection 13

5.2 Performance 15

6 Discussion **18**

7 Limitations **19**

8 Further Research **20**

9 Conclusion **20**

A Appendix **22**

Reference List **24**

List of Tables

1 Causal effects on long-term unemployment suggested by past literature. 4

2 Variable overview. 12

3 Selected features for logistic regression and SVM. 14

4 Predictor weights for logistic regression. 16

5 Performance measures for logistic regression and SVM. 17

6 Confusion matrix for the logistic regression model on training data. . . 22

7 Confusion matrix for the logistic regression model on test data. 22

8 Confusion matrix for the kernel SVM on training data. 23

9 Confusion matrix for the kernel SVM on test data. 23

List of Figures

1 Methodological procedure with sample sizes and number of independent variables. 10

2 Predictor coefficients for logistic regression depending on regularization parameter λ. 13

3 Predictor importance in SVM measured by AUC. 15

4 ROC curves on test data for logistic regression and kernel SVM. 17

List of Abbreviations

ANOVA	analysis of variance
AUC	area under the ROC curve
CV	cross-validation
ERM	empirical risk minimization
JSCI	jobseekers' classification instrument
lasso	least absolute shrinkage and selection operator
MCA	multiple component analysis
RKHS	reproducing kernel Hilbert space
ROC	receiver operating characteristic
SEM	structural equation modelling
SOEP	socio-economic panel
SVM	support vector machine
USA	United States of America

1 Introduction

Losing a job is an involuntary imprinting experience with far-reaching consequences for the affected people and their families. It is desirable to get a new job quickly but that is not always feasible. Some unemployed need specific state aid to escape unemployment. The purpose of this seminar thesis is to derive determinants for long-term unemployment in Germany and use them as predictors to classify unemployed individuals into classes of high and low risk of getting long-term unemployed. Furthermore, the thesis compares the performance of the in this application context widely used logistic regression approach and a support vector machine (SVM) with Gaussian kernel. This thesis expands the literature on economic forecasts and predictions with machine learning approaches. The model can be applied in the context of targeting for governmental unemployment assistance.

Job loss is psychologically damaging for individuals that experienced such an event. Mousteri et al. (2018) found that this damage is not only temporary but affects psychological and physical well-being up to old age. This lack in perceived quality of life in older age increases in unemployment duration. If the job loss took place in a period of weak economy, the decline in psychological and physical well-being is even worse (Urbanos-Garrido and Lopez-Valcarcel, 2015). Reemployment helps to mitigate this decline (Brand, 2015). Furthermore, there is a decline in career quality after an unemployment period which differs between male and female individuals (Manzoni and Mooi-Reci, 2020). Thus, prompt reemployment after job loss is desirable in the context of sustainable life quality and psychological well-being. Some individuals manage to get a new job quickly on their own but others need more time to find a new job or even will never get employed again. The result is long-term unemployment. In order to address the problem of long-term unemployment and associated lacks in well-being, a tool for long- and short-term unemployment classification is needed. Previous literature provides a multidisciplinary model to explore determinants of reemployment success (Wanberg et al., 2002). Furthermore, a prediction model for job-seekers' risk of staying unemployed more than 12 months has been developed by Matty (2013). Both models refer to a logistic regression approach. Kreiner and Duca (2020) and Katris (2020) showed that machine learning methods can outperform traditional economic models in forecasting unemployment rates. This seminar thesis refers to these findings and applies a kernel SVM to the long-term unemployment classification problem and compares its predictive performance with the logistic regression approach. It uses data from the German socio-economic panel (SOEP) to perform a feature selection through penalized logistic regression with a least absolute shrinkage and selection operator (lasso) and predictor importance measures for the SVM before training the final models.

The next section gives an overview about past publications and approaches to find causal effects on long-term unemployment as well as features to predict the risk of long-term unemployment. It motivates the choice of features used for the feature selection in this study. *Section 4* deals with the data used in this thesis followed by the description of methods applied. The results are presented in *Section 5* and discussed in *Section 6*. *Section 7* indicates limitations of the analyses. *Section 8* points out starting points for further research. Finally, *Section 9* concludes the thesis.

2 Context

There is a broad range of previous publications in analyzing job search behavior of unemployed people, reemployment success and long-term unemployment. These topics are related to each other in the sense that job search behavior, including search strategy, competencies, motives and constraints, has a significant impact on reemployment success (Wanberg et al., 1996). Wanberg et al. (2020) and van Hoye et al. (2009) investigated the importance of networking activities in job search behavior. Granovetter (1973) found that most new jobs come from networking through weak social ties. Studies that investigated predictors and effects on long-term unemployment are shown in *Table 1*. Note that duplicates in the column of effects have been removed.

Research on determinants of long-term unemployment is multidisciplinary field. Wanberg et al. (2002) considered predictor variables from economics, sociology and psychology to formulate a multidisciplinary model. Furthermore, the authors defined six job-seeker specific categories associated with reemployment success: job-search intensity and quality, economic need to work, reemployment constraints, social capital and human capital. Job-search intensity and quality captures all variables associated with time and effort an individual spent on job-searches including job seeking support (Wanberg et al., 1996). De Battisti et al. (2016) found that job-search intensity and reemployment success is determined by perceived employability and psychological distress. Variables representing an individual's financial resources during unemployment can be summarized in the group of economic need to work. This includes household income and number of children in household (Marelli and Vakuenko, 2016; Wanberg et al., 2002). The category of reemployment constraints represents determinants like access to public transports, health and willingness to move (O'Connell et al., 2012; Marelli and Vakuenko, 2016). Social capital refers to an individual's networking activities. Curtis et al. (2016) highlights the importance of in-group identification. The authors found that the employment status of friends and relatives influences reemployment success. Human capital refers to the abilities provided by the job seeker that could be transformed into employer's benefits including schooling and self-control. Besides these categories Krause (2013) found an effect of satisfaction on reemployment success. Additionally, psychological factors like depressive symptoms have an impact on reemployment (Kokko et al., 2000). *Section 4* will provide an overview about the specific variables used in this thesis.

Matty (2013) developed a model to predict the likelihood of long-term unemployment for job-seekers in the United Kingdom. This classification model can be used as an instrument to derive a JSCI-score that gives information about the risk for a specific individual to stay unemployed more than twelve months. The model refers to a logistic regression approach with feature selection through trying different combinations of predictors. The author decided to not take the most important but the most efficient features as predictors in the final model. The final model contains variables from the categories of individual's economic need to work an reemployment constraints as attribute variables. Attitudinal variables referring to the individual's agreement on a five point Likert scale with statements about motivation and confidence have been added. Finally, the model contains administrative predictors including the local average house price and proportion of working-age population employed in the local area. For calculating the JSCI-score the rounded and partly adjusted coefficient estimates serve as weights. The aim of this thesis is to take up the approach and consider the problem from a machine learning point of view. The logistic regression approach can be trans-

Publication	Country	Method	Effects
Wanberg et al. (1996)	USA	Logit	Age Job seeking support
Obben et al. (2002)	New Zealand	Logit	Gender Ethnicity Schooling Regional location
Marelli and Vakuenko (2016)	Russia, Italy	Probit	Marital status Health Household income
O'Connell et al. (2012)	Ireland	Probit	Employment history Willingness to move Unemployment duration Public transports
Curtis et al. (2016)	Australia	ANOVA	In-group identification
Kokko et al. (2000)	Finland	MCA	Depressive symptoms Self-control
Lötters et al. (2013)	Netherlands	Logit	Perceived Health Willingness to accept a job
Lallukka et al. (2019)	Finland	Logit	Social determinants
Krause (2013)	Germany	Logit	Satisfaction
Wanberg et al. (2002)	USA	Logit	Children in household Economic hardship
De Battisti et al. (2016)	Italy	SEM	Perceived employability Psychological distress

Table 1: Causal effects on long-term unemployment suggested by past literature.

formed into an empirical risk minimization (ERM) problem that is considered for the sake of comparison. The ERM framework allows to apply other powerful models, such as the the kernel SVM which will be explained in the next section.

3 Method

To address the classification task of predicting long-term unemployment previous studies mostly referred to logistic regression (Matty, 2013). Feature selection was either made by hand to investigate specific effects or by trying different combinations of features to detect efficient predictors. This thesis applies a kernel SVM and compares its performance to the widely used logistic regression approach. It performs feature selection from a machine learning point of view by applying lasso regularization and considering receiver operating characteristic (ROC) measures to evaluate the predictors' importance. In the first step a large number of possible predictors selected in *Section 4* is considered. After applying feature selection methods a logistic regression model and a kernel SVM is estimated based on the selected features and a training data set. The training data set contains 80% of the original number of observations. The withheld observations are used to evaluate and compare the models' performance in the last step. Machine learning techniques have recently been applied in unemployment context in time series analysis to forecast unemployment rates (Kreiner and Duca, 2020; Katris, 2020). The results showed a considerable improvement in forecasting performance compared to traditional econometric approaches.

The following subsections give insights into the applied models in this study and their respective feature selection procedure. The methods are based on the general ERM framework (Shalev-Schwartz and Ben-David, 2014, pp. 36ff.). Consider input space $\mathcal{X}$, actual output space $\mathcal{Y}_{act}$, prediction output space $\mathcal{Y}_{pred}$, centered and normalized training data points $(x_i, y_i)_{i=1,\dots,n} \in \mathcal{X} \times \mathcal{Y}_{act}$, function space $\mathcal{F} : \mathcal{X} \to \mathcal{Y}_{pred}$ and loss function $\ell : \mathcal{X} \times \mathcal{Y}_{act} \times \mathcal{Y}_{pred} \to \mathbb{R}_{\geq 0}$. In most cases it holds that $\mathcal{Y}_{pred} = \mathcal{Y}_{act}$ but in some cases they may differ as shown in *Section 3.1.* The function that leads to the smallest empirical risk is given by

$$f_n \coloneqq \operatorname*{argmin}_{f \in \mathcal{F}} R_n(f) \tag{1}$$

with empirical risk

$$R_n(f) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i, f(x_i)). \tag{2}$$

The framework can be extended to a regularized ERM framework by adding a regularizer $\Omega : \mathcal{F} \to \mathbb{R}_{\geq 0}$ with regularization constant λ to the empirical risk. The result is the regularized empirical risk

$$R_{reg,n}(f) \coloneqq R_n(f) + \lambda \, \Omega(f), \tag{3}$$

where λ is a hyperparameter.

3.1 Logistic Regression

The goal of this subsection is to introduce the logistic regression as an ERM problem (Shalev-Schwartz and Ben-David, 2014, pp. 126ff.). Furthermore, the technical point of view on the feature selection through lasso regularization will be presented.

Referring to the general ERM framework let d be the number of selected features including an intercept if needed, $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y}_{act} = \{\pm 1\}$ and $\mathcal{Y}_{pred} = (0, 1)$. Defining the logistic loss as

$$\ell_{log}(x_i, y_i, f(x_i)) \coloneqq log\big(1 + exp(-y_i f(x_i))\big) \tag{4}$$

and $\mathcal{F} = \{f(x) = \langle w, x\rangle;\ w \in \mathbb{R}^d\}$ leads to the logistic ERM problem

$$\underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} log\,(1 + exp\,(-y_i \langle w, x_i\rangle)). \tag{5}$$

A new data point $x_{new} \in \mathbb{R}^d$ can be classified using

$$P(x_{new}) = \frac{1}{1 + exp(-\langle w, x_{new}\rangle)}, \tag{6}$$

$$y^{label} = \begin{cases} 1 & \text{if} \quad P(x_{new}) \geq 0.5, \\ -1 & \text{if} \quad P(x_{new}) < 0.5. \end{cases} \tag{7}$$

Before estimating the coefficients w of this model a feature selection through lasso regularization will be applied (Tibshirani, 1996). Let D be the number of all considered features and m be the number of corresponding available training observations $(x_i, y_i)_{i=1,\dots,m} \in \mathcal{X} \times \mathcal{Y}_{act}$. Defining the lasso regularizer as

$$\Omega_{lasso}(f) \coloneqq ||w||_1 \tag{8}$$

leads to the regularized logistic ERM problem

$$\underset{w \in \mathbb{R}^D}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} log\,(1 + exp\,(-y_i \langle w, x_i\rangle)) + \lambda\,||w||_1. \tag{9}$$

The lasso regularizer punishes large coefficients in w and pulls them towards zero. In this process, the regularizer allows for some large coefficients if at the same time small coefficients shrink to zero (Tibshirani, 1996). Thus, an increasing hyperparameter λ leads to an increase of the number of coefficients that equal zero in the optimization. λ is set via five times repeated five-fold cross-validation (CV). In the application of these methods in this thesis it holds that $m \leq n$ due to the increased number of missing values when considering a larger number of features. Note that the weights w must not be interpreted causally. The aim in this application is to predict long-term unemployment as good as possible. The unregularized ERM method optimizes w, such that this goal is reached with respect to $\mathcal{F}$. Thus, w represents no causal effects but the feature weights that lead to the best cross-validated predictive power.

3.2 Support Vector Machine

The SVM is a powerful machine learning classification tool that can be extended by kernelization and can deal with high dimensional feature spaces (Cortes and Vapnik, 1995). Using the kernel trick the SVM can be applied non-linearly without too much overfitting and with only few tuning parameters. The purpose of this subsection is to define the SVM in the ERM framework and show how to kernelize this method. Feature selection is performed by evaluating the importance of each predictor in the model.

Referring to the general regularized ERM framework let d be the number of selected features including no intercept, $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y}_{act} = \{\pm 1\}$ and $\mathcal{Y}_{pred} = \mathbb{R}$. Defining the hinge loss as

$$\ell_{hinge}(x_i, y_i, f(x_i)) := max\,\{0; 1 - y_i\, f(x_i)\} \tag{10}$$

and regularizer

$$\Omega_{ridge}(f) := ||w||^2 \tag{11}$$

leads to the regularized ERM problem

$$\underset{w \,\in\, \mathbb{R}^d}{\operatorname{argmin}}\ \frac{1}{n}\sum_{i=1}^{n} max\,\{0; 1 - y_i\, f(x_i)\} + \lambda\, ||w||^2. \tag{12}$$

Rearranging constants in the optimization problem results in the ERM problem for SVM

$$\underset{w \,\in\, \mathbb{R}^d}{\operatorname{argmin}}\ \frac{C}{n}\sum_{i=1}^{n} max\,\{0; 1 - y_i\, f(x_i)\} + ||w||^2, \tag{13}$$

where C is a hyperparameter that behaves inversely to λ in (12) and will be set via five times repeated five-fold CV (Hastie et al., 2015, pp. 46ff.). When defining $\mathcal{F} = \{f(X) = \langle w, X\rangle + b;\ w \in \mathbb{R}^d, b \in \mathbb{R}\}$ the minimization problem can be transformed to the corresponding dual maximization problem

$$\max_{\alpha \in \mathbb{R}^n} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,\,j=1}^{n} \alpha_i\, \alpha_j\, y_i\, y_j\, \langle x_i,\, x_j \rangle \tag{14a}$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq \frac{C}{n} \tag{14b}$$

$$\sum_{i=1}^{n} \alpha_i\, y_i = 0. \tag{14c}$$

Since strong duality holds, both optimization problems lead to the same solution (Shalev-Schwartz and Ben-David, 2014, pp. 211f.). The dual problem has the important property that the input data only occurs in vector products. This property can be used to implicitly embed the data in higher and even an infinite number of dimensions through a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Explicit embedding would be computationally expensive but is not necessary here since single data points do not affect the model but only their vector products (Boser et al., 1992). A kernel function implicitly embeds the data into a reproducing kernel Hilbert space (RKHS). Considering the data in higher dimensions makes the SVM linearly applicable and leads to a non-linear hyperplane in $\mathbb{R}^d$. The Gaussian kernel is defined as

$$k_{gauss}(x_i, x_j) \coloneqq exp\Big(- \frac{||x_i - x_j||^2}{2\,\sigma} \Big), \tag{15}$$

where σ is a hyperparameter that gives information about the maximal distance where data points are treated as very similar in. Replacing the vector product by the implicit embedding through the Gaussian kernel leads to the optimization problem

$$\max_{\alpha \in \mathbb{R}^n} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,\,j=1}^{n} \alpha_i\, \alpha_j\, y_i\, y_j\, exp\Big(- \frac{||x_i - x_j||^2}{2\,\sigma} \Big) \tag{16a}$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq \frac{C}{n} \tag{16b}$$

$$\sum_{i=1}^{n} \alpha_i\, y_i = 0, \tag{16c}$$

where C and σ are hyperparameters that will be set via five times repeated five-fold CV. Assuming the optimization problem has been solved and defining $J = \{j \,|\, 0 < \alpha_j < \frac{C}{n}\}$, i.e. any point x_j is a support vector, the function value for new data point $x_{new} \in \mathbb{R}^d$ can be calculated as

$$f(x_{new}) = \sum_{i=1}^{n} \alpha_i\, y_i\, k(x_i,\, x_{new}) + \Big(y_j - \sum_{i=1}^{n} y_i\, \alpha_i\, k(x_i,\, x_j) \Big); \quad j \in J. \tag{17}$$

The class decision is made based on the function value

$$y^{label} = \begin{cases} 1 & \text{if} \quad f(x_{new}) \geq 0, \\ -1 & \text{if} \quad f(x_{new}) < 0. \end{cases} \tag{18}$$

The lasso feature selection method cannot be applied here. Therefore, feature selection will be performed through backward elimination (Guyon and Elisseeff, 2003). First, the model is optimized with the full set of variables. Afterwards the least important features measured by area under the ROC curve (AUC) will be removed for the final model.

3.3 Model Evaluation

After the model training the models are evaluated according to their predictive power by using two measures: the risk based on the 0-1-loss and the AUC. The risk summarizes the relative amount of data points that are wrongly classified by a model. Defining the 0-1-loss as

$$\ell_{0-1}(x_i, y_i, y^{label}) := \begin{cases} 0 & \text{if} \quad y_i = y^{label}, \\ 1 & \text{if} \quad y_i \neq y^{label} \end{cases} \tag{19}$$

the empirical risk is given by (2). The risk approach can be applied to the models' performance on the training set (empirical risk), during the CV and on the test set. For the CV risk the average risk across folds is calculated.

Another widely used classifier performance measure is the AUC. It is a single number measure that gives information about the performance of a classifier at different decision thresholds. At the same time the AUC is independent from a priori class probabilities and thus, comparable across studies (Bradley, 1997). The ROC curve plots the model's true positive rate depending on the false positive rate. Thus, an AUC value equal to 0.5 corresponds to a model that performs as good as a random classifier whereas an AUC value greater than 0.7 is associated with acceptable discrimination (Hosmer et al., 2013, pp. 173ff.).

4 Data

The analyses and models in this seminar thesis are estimated using data from the German SOEP. It contains annual multidisciplinary survey data about approximately 30,000 individuals in 15,000 households (diw.de, accessed 04-January-2021). The SOEP provides information for many of the determinants of long-term unemployment suggested by previous literature presented in *Section 2*. Since 2009 it seamlessly records individuals' unemployment history on a monthly basis that allows unambiguous labeling in short- and long-term unemployment. Individuals are considered if there has been at least one period of unemployment that is either not censored from the right or already lasted longer than twelve months and if the individual has been participated in the survey at least once during that period. If there is more than one period for a specific individual that fulfills the criteria, the most recent one is taken. If an individual participated more than once during the unemployment period of interest, only data from the first interview during that period makes it in the data set. Accordingly, every individual is associated with a unique observation in the data set. *Table 2* gives an overview about the variables considered in the first methodological step before feature selection.

The data set contains 2,524 labeled observations. Due to missing values the data set shrinks depending on the number of variables used in the respective methodological step. The relative amount of long-term unemployed individuals is barely affected by

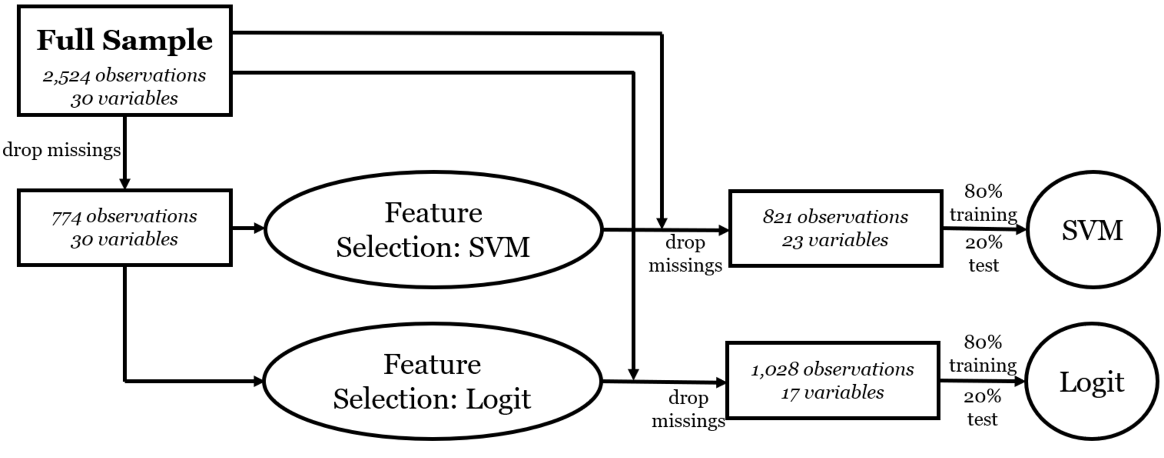

Figure 1: Methodological procedure with sample sizes and number of independent variables.

this selection process. The methodological procedure as well as the corresponding number of observations and independent variables are shown in *Figure 1.* All variables from *Table 2* are considered during the process of feature selection. The sample size reduces to 774 observations with 54.78% long-term unemployment rate. After feature selection for the logistic regression approach the sample contains 1,028 individuals, where 53.89% of them are long-term unemployed. Accordingly, for SVM the sample consists of 821 observations with 54.81% long-term unemployment rate.

Some variables are not part of every annual wave of the SOEP survey. For those variables forward or backward imputation is applied, depending on the characteristic of the variable. For variables referring to character traits and individual's assessment of other people the temporally closest data point is considered. For variables concerning the social network the most recent data point is contemplated for imputation.

Variable	**Description**	**Coding**
unemployed_long	Long-term unemployed	dummy
age	Age	integer
gender	Gender	dummy
german	German nationality	dummy
marital_status	Marital status: married, living together married, separated single divorced widowed registered same-sex partnership	factor
children	Number of children living in household: 0 1 2 3 > 3	factor
hh_income	Household net income	integer
worr_crime	Worried about crime in Germany	3-point Likert-Scale
worr_econ	Worried about economic development	3-point Likert-Scale
worr_finances	Worried about finances	3-point Likert-Scale
sat_dwelling	Satisfaction with dwelling	11-point Likert-Scale

sat_pincome	Satisfaction with personal income	11-point Likert-Scale
sat_hincome	Satisfaction with household income	11-point Likert-Scale
sat_life	Satisfaction with life at present	11-point Likert-Scale
friends	Amount of closed friends: < 6 6-10 11-15 16-20 > 20	factor
health	Current health	5-point Likert-Scale
disabled	Severely disabled	dummy
empl_interest	Employment interest: full-time part-time both not sure	factor
caution_foreigners	Caution towards foreigners	4-point Likert-Scale
suitable_pos	Difficulty to find suitable position.	3-point Likert-Scale
empl_intended	Employment intended	4-point Likert-Scale
acceptance	Could start working within 2 weeks	dummy
state_exploitive	Agree: Most people are exploitive.	dummy
state_helpful	Agree: People usually try to be helpful.	dummy
state_trust	Agree: People can generally be trusted.	4-point Likert-Scale
state_losers	Agree: Most people are basically losers.	6-point Likert-Scale
state_special	Agree: Being a very special person gives me a lot of strength.	6-point Likert-Scale
state_personality	Agree: I deserve to be seen as a great personality.	6-point Likert-Scale
state_show	Agree: I react with annoyance if another person steals the show from me..	6-point Likert-Scale
state_rivals	Agree: I want my rivals to fail.	6-point Likert-Scale

Table 2: Variable overview.

5 Results

According to the methodological procedure presented in *Section 3* and illustrated in *Figure 1* the following subsections show and explain the results of each step. The first subsection deals with the outcomes of the feature selections through penalized logistic regression as well as variable specific AUC in SVM and compares them with each other. The second subsection presents the performances of the final models.

5.1 Feature Selection

Two different methods for feature selection have to be applied in the context of logistic regression and kernel SVM. First, this subsection presents the results of the feature selection for logistic regression through lasso penalization. The coefficients of all predictors are pulled towards zero in increasing regularization parameter λ. *Figure 2* shows the predictor selection process depending on λ. The eight most important features are colored. Additionally, the optimal $\lambda = 0.0164$ found by CV that lead to the smallest CV risk is highlighted in red,

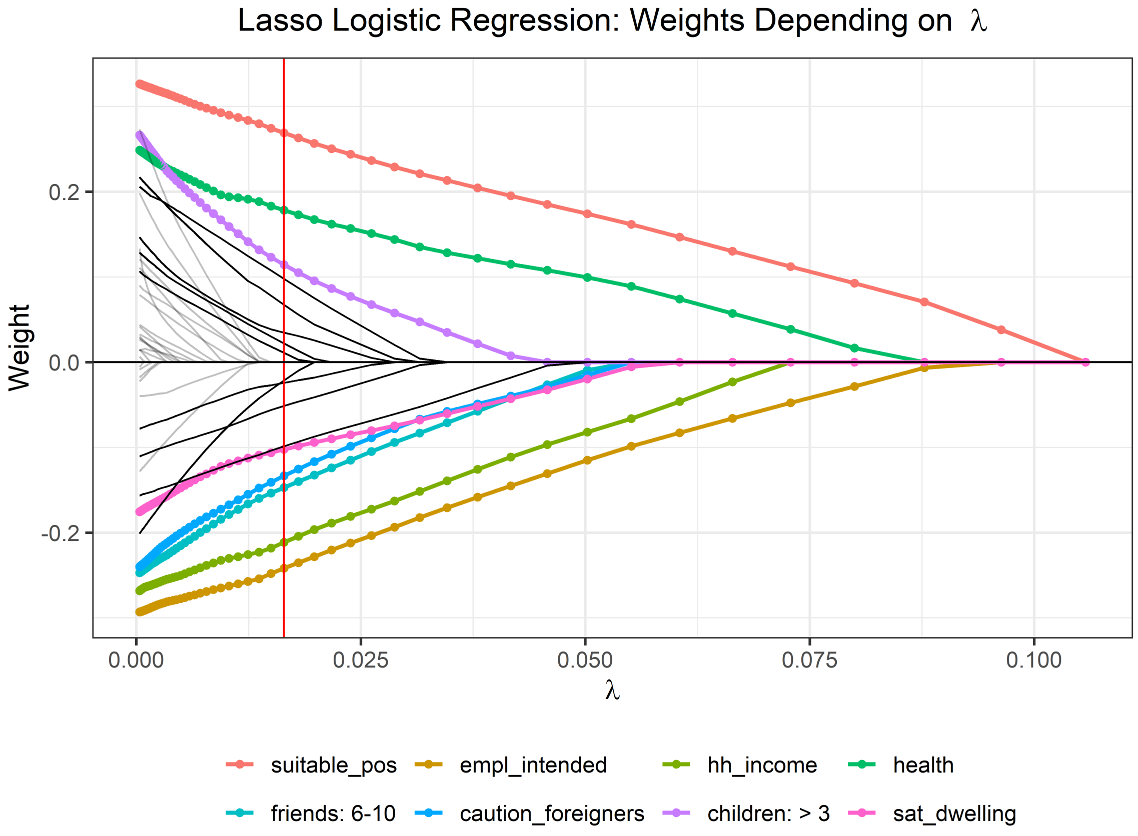

Figure 2: Predictor coefficients for logistic regression depending on regularization parameter λ.

such that the weights for the predictors that have not been selected turn zero for values of λ smaller than 0.0164. Besides the fact that λ is cross-validated, the graph also shows a small gap between the value for λ at which the last feature that has not been selected drops out and the value at which the first feature that has been selected shrinks to zero. The most important predictors include difficulty to find a suitable position, household net income, health and number of children living in the same household as the respondent. This coincides with the results of Matty (2013). All selected features for logistic regression are listed in *Table 3*.

In the case of kernel SVM variable specific importance is considered for the feature selection process. When cross-validating the model hyperparameters C and σ the ROC and corresponding AUC for each parameter is calculated and used as an importance measure. The CV with all considered features results in $C = 0.7407$ and $\sigma = 0.0324$. The corresponding importance for the 30 most important features are shown in *Figure 3*. After trying different feature settings inspired by backward elimination, the best predictive power could have been reached when selecting features with at least an importance of 0.52 measured by the AUC. The threshold is highlighted in red in *Figure 3*. The most important features such as difficulty to find a suitable position, household income and health coincide with the results from logistic regression and the findings of Matty (2013). Nevertheless, there are some differences between the selected features for logistic regression and SVM.

<table>
<tr><th>Logistic Regression</th><th>SVM</th></tr>
<tr><td colspan="2"><i>worr_crime</i></td></tr>
<tr><td colspan="2"><i>sat_dwelling</i></td></tr>
<tr><td colspan="2"><i>health</i></td></tr>
<tr><td colspan="2"><i>caution_foreigners</i></td></tr>
<tr><td colspan="2"><i>friends</i></td></tr>
<tr><td colspan="2"><i>marital_status</i></td></tr>
<tr><td colspan="2"><i>acceptance</i></td></tr>
<tr><td colspan="2"><i>empl_intended</i></td></tr>
<tr><td colspan="2"><i>state_personality</i></td></tr>
<tr><td colspan="2"><i>hh_income</i></td></tr>
<tr><td colspan="2"><i>suitable_pos</i></td></tr>
<tr><td colspan="2"><i>empl_interest</i></td></tr>
<tr><td><i>german</i></td><td><i>worr_econ</i></td></tr>
<tr><td><i>children</i></td><td><i>worr_finances</i></td></tr>
<tr><td><i>state_helpful</i></td><td><i>sat_hincome</i></td></tr>
<tr><td></td><td><i>sat_life</i></td></tr>
<tr><td></td><td><i>gender</i></td></tr>
<tr><td></td><td><i>severely_disabled</i></td></tr>
<tr><td></td><td><i>age</i></td></tr>
<tr><td></td><td><i>state_losers</i></td></tr>
<tr><td></td><td><i>state_trust</i></td></tr>
<tr><td></td><td><i>state_special</i></td></tr>
</table>

Table 3: Selected features for logistic regression and SVM.

Table 3 shows a full comparison of selected features for both approaches. It is conspicuous that the feature selection for SVM results in considerably more selected features in comparison to the penalized logistic regression. The features only selected by kernel SVM are specific worries and satisfaction as well as individual's assessment of other people. Furthermore, the

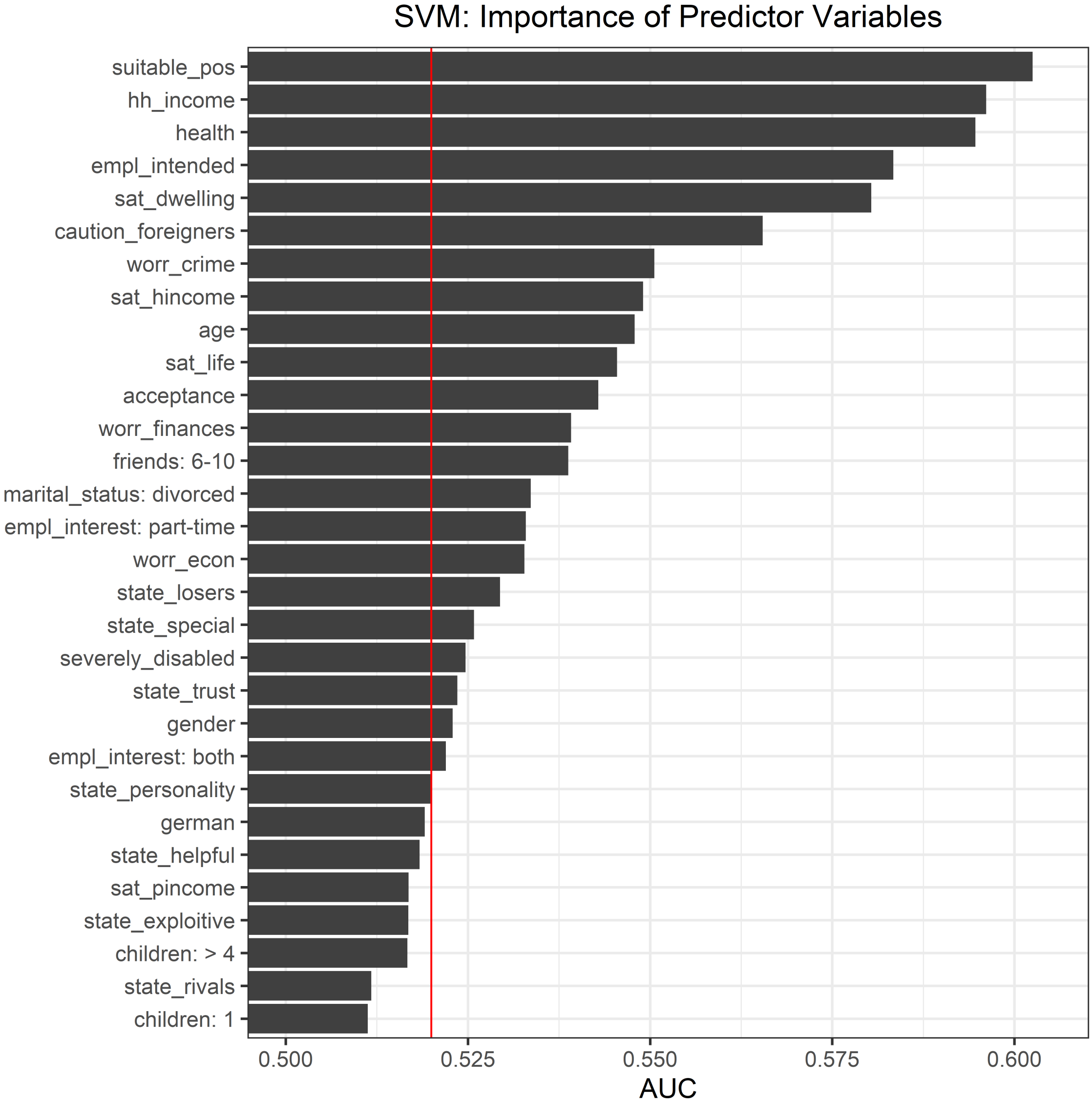

Figure 3: Predictor importance in SVM measured by AUC.

significant overlap of features is noticeable. Only three predictors that have been selected by penalized logistic regression have not been selected by the SVM variable importance selection procedure.

5.2 Performance

After detecting relevant features for both approaches the final models are trained. The model parameters are set through five times repeated five-fold CV on 80% of the observations in the model specific samples. The hyperparameters $C = 0.4383$ and $\sigma = 0.0165$ lead to the smallest CV risk for kernel SVM. The predictor weights for the logistic approach are shown in *Table 4*. Note that in this case of application the weights should neither be interpreted

Predictor	**Weight**
intercept	0.368
worr_crime	-0.1371
sat_dwelling	0.0067
health	0.2491
caution_foreigners	-0.1354
suitable_pos	0.3268
empl_interest: part-time	0.3594
empl_interest: both	0.4311
empl_interest: not sure	-0.0492
friends: 6-10	-0.3279
friends: 11-15	0.4557
friends: 16-20	0.4473
friends: > 20	-0.7001
marital_status: married, separated	-0.1879
marital_status: single	0.3402
marital_status: divorced	0.3495
marital_status: widowed	0.1948
marital_status: registered same-sex partnership	-12.5423
german	-0.6888
acceptance	-0.2215
state_helpful	0.3678
empl_intended	-0.3356
children: 1	-0.0082
children: 2	0.2145
children: 3	0.0831
children: > 3	1.2176
state_personality	0.1346
state_rivals	-0.0773
hh_income	-0.4097

Table 4: Predictor weights for logistic regression.

regarding causality nor with respect to significance (see *Section 3.1*). The marital status of being registered in a same-sex partnership can be used as an example here. Since there are only few individuals in the sample having this characteristic, the model is likely to overfit and the weight cannot be interpreted as causal.

Table 5 shows the performance measures for the logistic regression model and the kernel SVM. Note, that the CV risk for logistic regression is simulated because originally CV is not necessary since no hyperparameter is included. The epirical risks for logistic regression are smaller than those for kernel SVM. The baseline accuracy is 46.11% for the logistic regression sample and 45.19% for kernel SVM respectively. Thus, the test risk of both approaches is smaller than the baseline. According to the AUC, the logistic regression performs better on

Measure	Logistic Regression	SVM
Empirical risk	0.3147	0.352
CV risk	0.3573	0.399
Test risk	0.3527	0.3879
Training AUC	0.74	0.7157
Test AUC	0.6628	0.732

Table 5: Performance measures for logistic regression and SVM.

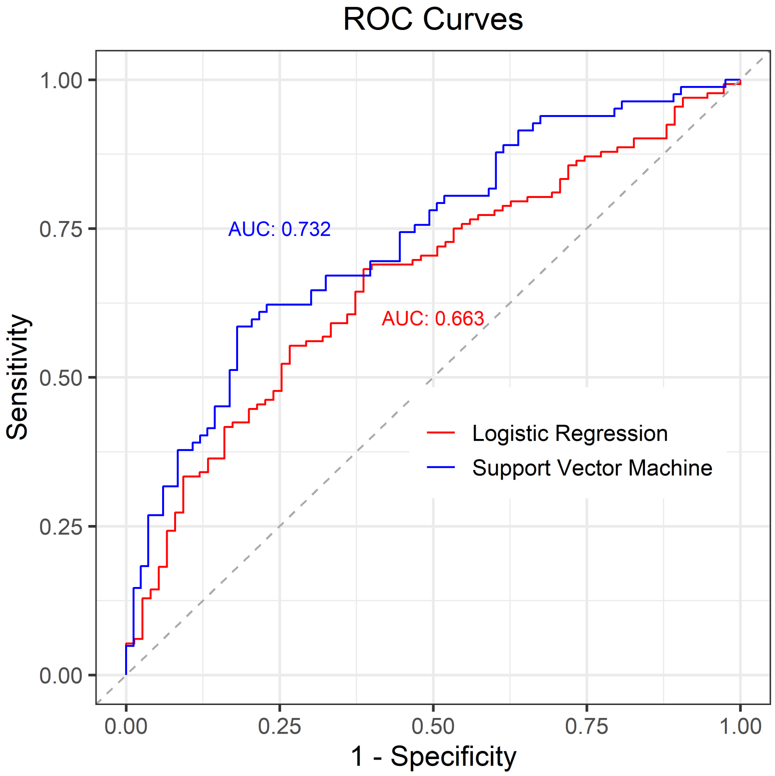

Figure 4: ROC curves on test data for logistic regression and kernel SVM.

the training data than the kernel SVM. In terms of the test AUC the kernel SVM yields better results on test data. The test AUC for the SVM model is 0.732 which is associated with an acceptable discrimination following the taxonomy of Hosmer et al. (2013). In contrast, the test AUC of the logistic regression approach misses the threshold of 0.7. In consequence, the test AUC of 0.6628 is considered as poor discrimination. The corresponding ROC curves are illustrated in *Figure 4*, where the dashed grey line serves as a reference curve. The distance between the ROC curve for kernel SVM and the reference is larger than the curve for logistic regression. For nearly any decision threshold the SVM is considered as the more accurate model. The appendix includes confusion matrices for more insights.

6 Discussion

The results showed a significant overlap of the predictors for logistic regression and the kernel SVM. 12 of 15 predictors that have been selected through penalized logistic regression are also considered as important features for the kernel SVM. A comparison between the logistic regression approach and the results of Matty (2013) shows that the selected features agree to a great extend. The model developed by Matty (2013) found different factors that determine difficulty to find a job, such as lack of experience, lack of references as well as problems with public transports. For the logistic model in this thesis these factors are summarized in one measure for perceived difficulty to find a job. Furthermore, Matty (2013) found the number of children in the household, marital status and measures for self-confidence as important predictors that also appear in the selection of this study. Beyond that, the results of this thesis detected worries and degree of satisfaction as predictive features.

The feature selection using the AUC as an importance measure in kernel SVM resulted in considerably larger number of predictors. This can be explained by the induced limitations in the function space of logistic regression. Using the kernel trick in SVM makes the linear function space implicitly applicable in higher dimensions that enables this method to find similarities and dependencies that could not be identified by linear approaches in the original feature space. In the application case of this study the kernel SVM detected additional psychological predictors such as specific worries, character traits and life satisfaction.

Overall, the results of the feature selection show that predicting long-term unemployment requires multidisciplinary data and analyses and thus, agrees with the findings of Wanberg et al. (2002). The analysis revealed sociological, psychological and economic features as important predictors for long-term unemployment.

For comparing the performance of the logistic regression model and the kernel SVM two performance measures are taken into account. On the one hand, the logistic regression resulted in a smaller test risk compared with the test risk of SVM. On the other hand, SVM outperforms the logistic regression model when contemplating the AUC. *Figure 4* evidently shows better predictive power for the SVM measured by the AUC. Bradley (1997) highlighted important favorable properties of the AUC as a performance measure. The AUC is independent from the decision threshold and invariant to a priori class probabilities. For this reason the AUC measure is frequently used in studies for evaluating predictive performance. Matty (2013) also used this measure for the underlying logistic model of the JSCI-score. The model yielded an AUC of 0.7951 and thus, performs better than the comparable logistic approach of this thesis and the kernel SVM. This is either caused by fewer information in the selected data or by differences between countries. Matty (2013) had the opportunity to use geographical data and designed a specific survey for the selected sample of individuals. For measuring the difficulty for an individual to find a suitable position the author considered different items, whereas in this study the difficulty is only quantified by one objective measure. Nevertheless, the results of this thesis indicate that a kernel SVM can outperform the widely used logistic regression approach in predictive power. Furthermore, the results make it seem promising to approach this and similar prediction problems from a machine learning point of view.

7 Limitations

The results and findings of this thesis are based on the German SOEP data. Due to the combination of a large number of considered features and a not negligible missing rate, the number of workable observations shrinks significantly. A method to avoid this issue would be imputation based on similar data points. Since this can lead to more noise in the data this technique found no application in the analyses. The SVM approach can deal with the issue as a large margin classifier that fits a hyperplane only based on data points that serve as support vectors (Shalev-Schwartz and Ben-David, 2014, pp. 202ff.). The logistic regression approach can handle this problem since there is no systematic dropout of data points.

Since the SOEP data is anonymized it is not possible to extend the data with additional individual specific data. Thus, the analyses in this study are fully based on the limited range of variables included in the SOEP. However, the number of variables and the content of this limited range were sufficient to investigate the research problem of this thesis, even though it was not possible to reproduce a model with the same or higher predictive power than the model developed by Matty (2013).

In the methodological process of this thesis the AUC is used as an importance measure in the feature selection process and as a performance measure for model comparison. Although, this measure is widely used and recommended by Bradley (1997), some critics have arisen in recent years. Muschelli (2020) highlighted that the AUC measure may be misleading when applying it on binary or categorical variables. This concern mainly affects the application of the measure in feature selection context where it may overestimate the importance of features. Since the feature selection process also included backward elimination, the overall result is negligibly affected by this issue.

In comparison to Matty (2013) the results show differences in predictive power. This is either due to country-specific differences or due to quality of selected variables. Thus, the results may not be generalizable to other countries. This issue may be considered in further research by applying similar methods to data from other countries.

8 Further Research

The results of this thesis indicate that machine learning approaches can lead to better long-term unemployment predictions than traditional econometric approaches. The study focused on a kernel SVM in an ERM framework. Although, SVM in combination with the kernel trick is a powerful classification model, further research may consider other machine learning approaches. Moreover, no cost weights have been included in the models of this study. Depending on the application case it can be useful to adjust the loss function of the model, such that a misclassification for one specific class influences the value of the function more than a misclassification for the other class and thereby induce a cost-sensitive classification. Further research may include ethical considerations and psychological consequences. For the prediction through statistical models always includes a particular risk, psychological effects for individuals being classified as long-term unemployed may be investigated. Even the job-search behavior and thus reemployment success may be affected by knowing about own classification.

9 Conclusion

Past research found that unemployment and especially long-term unemployment has a long-lasting negative effect on life quality. Therefore, in favor of overall life quality it is a subject of interest to avoid continual unemployment. Moreover, preceding research put much effort in finding causal effects of individual's sociological, psychological and economic characteristics on reemployment success, unemployment duration and long-term unemployment. The research have been used by Matty (2013) to estimate a logistic regression prediction model and develop a JSCI-score for predicting unemployed individuals' risk of becoming long-term unemployed. This thesis used a systematic feature selection based on features suggested by preceding literature and showed that the classification problem of predicting long-term unemployment can be more successfully approached by machine learning methods. The study used the ERM framework and German SOEP data to estimate a baseline logistic regression model and a kernel SVM. For feature selection a lasso logistic regression and backward elimination based on feature importance has been applied. The results showed that the kernel SVM lead to better prediction results considering the AUC as a performance measure. Though, the SVM approach is not very suitable to develop a scoring scheme as the JSCI-score, but the ERM framework enables the user to induce cost-sensitiveness and to estimate classification probabilities.

In previous economic research machine learning has been applied in time series analysis and forecasting. This seminar thesis broadens the scope of machine learning for prediction in economic research. It shows that powerful methods like kernel SVM can outperform traditional models such as logistic regression in economic classification tasks. It encourages researchers to investigate further application cases.

A Appendix

Confusion Tables

Logistic Regression

actual \ predicted	short-term	long-term	Row Total
short-term	264	136	400
long-term	123	300	423
Column Total	387	436	823

Table 6: Confusion matrix for the logistic regression model on training data.

actual \ predicted	short-term	long-term	Row Total
short-term	46	29	75
long-term	44	88	132
Column Total	90	117	207

Table 7: Confusion matrix for the logistic regression model on test data.

Kernel SVM

actual \ predicted	short-term	long-term	Row Total
short-term	101	188	289
long-term	44	326	370
Column Total	145	514	659

Table 8: Confusion matrix for the kernel SVM on training data.

actual \ predicted	short-term	long-term	Row Total
short-term	33	50	83
long-term	14	68	82
Column Total	47	118	165

Table 9: Confusion matrix for the kernel SVM on test data.

Reference List

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *In 'Conference on Learning Theory (COLT)'*, 144–152.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

Brand, J. E. (2015). The far-reaching impact of job loss and unemployment. *Annual Review of Sociology*, *41*, 359–375.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Curtis, E., Gibbon, P., & Katsikitis, M. (2016). Group identity and readiness to change unemployment status. *Journal of Employment Counseling*, *53*, 50–59.

De Battisti, F., Gilardi, S., Guglielmetti, C., & Siletti, E. (2016). Perceived employability and reemployment: Do job search strategies and psychological distress matter? *Journal of Occupational and Organizational Psychology*, *89*, 813–833.

diw.de. (accessed 04-January-2021). https://www.diw.de/en/diw_01.c.615551.en/research_infrastructure__socio-economic_panel__soep.html

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations.* CRC Press, Taylor & Francis Group. https://web.stanford.edu/~hastie/StatLearnSparsity/index.html

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3^{rd}). Wiley.

Katris, C. (2020). Prediction of unemployment rates with time series and machine learning techniques. *Computational Economics*, *55*, 673–706.

Kokko, K., Pulkkinen, L., & Puustinen, M. (2000). Selection into long-term unemployment and its psychological consequences. *International Journal of Behavioral Development*, *24*(3), 310–320.

Krause, A. (2013). Don't worry, be happy? Happiness and reemployment. *Journal of Economic Behavior & Organization*, *96*, 1–20.

Kreiner, A., & Duca, J. (2020). Can machine learning on economic data better forecast the unemployment rate. *Applied Economics Letters*, *27*(17), 1434–1437.

Lallukka, T., Kerkelä, M., Ristikari, T., Merikukka, M., Hiilamo, H., Virtanen, M., Overland, S., Gissler, M., & Halonen, J. I. (2019). Determinants of long-term unemployment in early adulthood: A Finnish birth cohort study. *SSM - Population Health*, *8*, 100410.

Lötters, F., Carlier, B., Bakker, B., Borgers, N., Schuring, M., & Burdorf, A. (2013). The influence of perceived health on labour participation among long term unemployed. *Journal of Occupational Rehabilitation*, *23*, 300–308.

Manzoni, A., & Mooi-Reci, I. (2020). The cumulative disadvantage of unemployment: Longitudinal evidence across gender and age at first unemployment in Germany. *PLoS ONE*, *15*(6), e0234786.

Marelli, E., & Vakuenko, E. (2016). Youth unemployment in Italy and Russia: Aggregate trends and individual determinants. *The Economic and Labour Realtions Review*, *27*(3), 387–405.

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

Mousteri, V., Daly, M., & Delaney, L. (2018). The scarring effect of unemployment on psychological well-being across Europe. *Social Science Research*, *72*, 146–169.

Muschelli, J. (2020). ROC and AUC with a binary predictor: A potentially misleading metric. *Journal of Classification*, *37*, 696–708.

Obben, J., Engelbrecht, H.-J., & Thompson, W. (2002). A logit model of the incidence of long-term unemployment. *Applied Economics Letters*, *9*(1), 43–46.

O'Connell, P. J., McGuinness, S., & Kelly, E. (2012). The transition from short- to long-term unemployment: A statistical profiling model for Ireland. *The Economic and Social Review*, *43*(1), 135–164.

Shalev-Schwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms.* Cambridge University Press. http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Urbanos-Garrido, R. M., & Lopez-Valcarcel, B. G. (2015). The influence of the economic crisis on the association between unemploymenr and health: An empirical analysis for Spain. *European Journal of Health Economics*, *16*, 175–184.

van Hoye, G., van Hooft, E. A. J., & Lievens, F. (2009). Networking as a job search behaviour: A social network perspective. *Journal of Occupational and Organizational Psychology*, *82*(3), 661–682.

Wanberg, C. R., Hough, L. M., & Song, Z. (2002). Predictive validity of a multidisciplinary model of reemployment success. *Journal of Applied Psychology*, *87*(6), 1100–1120.

Wanberg, C. R., van Hooft, E. A. J., Liu, S., & Csillag, B. (2020). Can job seekers achieve more through networking? The role of networking intensity, self-efficacy, and proximal benefits. *Personnel Psychology*, *73*, 559–585.

Wanberg, C. R., Watt, J. D., & Rumsey, D. J. (1996). Individuals without jobs: An empirical study of job-seeking behavior and reemployment. *Journal of Applied Psychology*, *81*(1), 76–87.