Online job vacancy attractiveness: A tool to improve views, reactions and conversions.

Zuzana Košťalová[a], Štefan Lyócsa[a,b,c], Miroslav Štefánik[a,*]

[a]*Institute of Economic Research, Slovak Academy of Sciences, Sancova 56, 811 05 Bratislava, Slovakia*
[b]*Institute of Financial Complex Systems, Masaryk University, Lipova 41a, 602 00 Brno, Czech Republic*
[c]*Faculty of Management, University of Presov, Konstantinova 16, 080 01 Presov, Slovakia*

Abstract

The development of e-commerce has not escaped the labor market either, and employers seeking new labor force are posting vacancies on specialized web portals. Operators of such web portals and employers are interested in increasing the attractiveness of online job vacancies. However, given the extremely heterogeneous job types and job-seekers, it is difficult for employers and specialized web portal operators to design job offers that will lead to higher views, reactions and conversions (the ratio of the two). In a collaboration with a leading platform for online job vacancies in Slovakia, we study, whether machine learning methods can improve predictions of online job vacancies attractiveness on a sample of 32482 online job vacancies that offer as much as 883 job features. Our study show, that as opposed to various linear models, considerable prediction improvements can be achieved using the random forest. Based on this insight, we perform a statistical evaluation of key variables of importance. We find, that job classification, job benefits, and variables related to a simple morphological description of the job and job's title are relevant. Results of this study can help operators of specialized job vacancy portals and employers, to improve their job offers in order to attract more job seekers.

Keywords: online job vacancy; web-page attractiveness; web-page views; machine learning; e-recruitment, morphological text analysis.

Disclosure statement: No potential conflict of interest was reported by the authors. The authors are responsible for any errors.

Funding: This research is a result of the Diskow project, supported by the European Commission under the Erasmus+ programme: http://www.diskow.eu.

[*] Corresponding author

1. Introduction

In order to attract as many relevant job-seekers as possible, employers post vacancies. This behavior led to a design of many job searching models (see Rogerson et al. [2005] for a review), that describe an interaction between job-seekers and job-posting employers, under different labor market conditions. An important characteristic of such models is the job-finding rate that describes the speed at which job-seekers (e.g. unemployed) can secure a job. While most of job searching models seems to work with job-finding rates (see the baseline model of Pissarides [2000]), as noted by Davis et al. [2013], the counterpart, the job-filling rate, received less attention. Both are important characteristics of the labor market as higher job-finding and -filling rates suggest a faster convergence toward market equilibrium and in that sense a more efficient labor market. However, research related to job-finding is targeted more towards the behavior and characteristics of job-seekers, while under-researched job-filling on job posting employers. In the current era, online job vacancies (OJV) represent the predominant platform, where a matching between job supply and demand occur. Job-filling rate is crucial for employers aiming to fill-inn an open vacancy, and also for operators of web-based job market platforms.

Our study explores a relatively unexplored area related to job-filling literature, as well as to the emerging literature on the usefulness of data from OJVs (e.g. job recommender systems). Specifically, we are interested in two research questions. First, using data from individual online job vacancies, we are interested whether it is possible to improve attractiveness of an online job vacancy, that we measure as: i) number of *views*, ii) *reactions* (filling out an application form) and iii) *conversions* (the ratio of the later to former). As more views, reactions and conversions suggest a higher interest in a specific job vacancy, the three characteristics are related to the job-filling rate. Our potential set of explanatory variables initially consists of a large set of 883 variables that are grouped into 13 job description characteristic (e.g. job benefits, business sector, calendar effect, etc.). In order to assess the predictability of attractiveness of online job vacancies, we rely on standard machine learning technique and models, such as OLS, LASSO, Ridge, Elastic Net or random forest. Predictability and the extent of improved predictability would suggest that the job-filling rate, an important parameter in labor market models, can be improved by employers and job market platform. *Second*, we are interested in whether all groups of variables tend to predict attractiveness of OJV in the same way, i.e. if some variables tend to be more useful than others.

In the following section, we provide an overview of relevant studies that rely on OJV data. The third section describes our data-source, key variables of interest, namely views, reactions and conversions, as well as different categories of potential drivers of the attractiveness of online job vacancies. In this section, we also present data cleaning procedures. In the fourth section, we outline estimation techniques, namely regularization methods and random forest, forecast evaluation and a related statistical approach to variable importance that is able to identify which groups of explanatory variables are useful in predicting OJV's attractiveness. In the fifth section, we present our key results and the final, the sixth section, follows with our concluding remarks that also include alleys for future research.

2. Related literature

Rich data from online job vacancies (OJV) present a tempting source of information on various aspects of the labor market. Exploration of such data faces multiple challenges, e.g. representatives of the data sources, complexity associated with big data, text-based job descriptions leading to the need of natural language processing [Kureková et al., 2015], or handling of missing values and outliers. Recent advancement of techniques capable of addressing such challenges, have enabled a new generation of empirical studies, that are characterized by utilizing data from individual OJVs, to be a complement or alternative of aggregate (un)employment or labor market survey data (e.g. Fabo et al. 2017, Turrell et al. 2018a, Lovaglio et al. 2020, Nicole et al. 2020).

Online job vacancy market is dominated by commercial providers [Cedefop, 2019], therefore data are rarely publicly available. Yet, studies using OJV data continue to add up (e.g. [Fabo et al., 2017, Turrell et al., 2018a, Deming and Kahn, 2018][1]). De Pedraza et al. [2019], Lovaglio et al. [2020] have provided evidence that OJV data has the potential to complement or substitute the official vacancy statistics. Turrell et al. [2018b] use OJV data from the UK to identify the effect of job mismatch on firm-level output and productivity. Marinescu and Wolthoff [2020] attempt to identify job posting characteristics that explain cross-section variances in posted wages, required education and experience. Interestingly, job titles appears to be the most relevant driver overall. Apart from such studies, three early streams in the literature can be identified.

2.1. Job classification

Turrell et al. [2018a] adopt an empirically driven bottom-up approach to analyse labour market segmentation on a data set of OJV from the UK between 2008-2016. Applying machine-learning text analysis, they cluster vacancies based on their content and compare yielded clusters to categories of established occupational classifications. This approach allows them to identify

[1] For an overview of seminal studies, see Kureková et al. [2015].

"new careers" and explore clusters' power to discriminate between offered job salaries. Boselli et al. [2018] propose the so called WoLMIS system that relies on OJV data to perform job classification tasks. The empirical part is based on OJV data from UK and Ireland. The WoLMIS system is used to classify job postings and the accuracy is compared against ISCO classification codes that were given to jobs by domain experts. Colace et al. [2019] has used a similar data as Boselli et al. [2018] and also relied on machine-learning system to classify job posting into eight information communication technology (ICT) related professions (according to ISCO 4 digit classification). Our research is related to this stream of the literature in that we also explore the role of job classification, specifically for the attractiveness of online job vacancies.

2.2. Skill classification

Using data from Italian OJV, Lovaglio et al. [2018] and Colombo et al. [2019] identified different types of skills required by by employers. In Lovaglio et al. [2018] the interest was to identify skills that best discriminate between statisticians from other internet communication technology ICT jobs. Deming and Kahn [2018] explore a firm-level variance in skill requirements and show the importance of within-occupational skills differences observed in OJVs in explaining wage differences, as well as firm performance. Grinis [2019] shows how the UK employers skills requirements of Science, Technology, Engineering and Mathematics (STEM) spill over between STEM and non-STEM occupations. These results suggest, that a proper specification of a statistical model that predicts attractiveness of online job vacancies should also include skills (see Section 3). Recently, also Stephany [2020] was interested in how learning a new skill from a different field might increase worker's wages. He used data from job profiles of freelancers using a crowd-sourcing platform and found evidence on a benefit related to cross-skilling, but results are heterogeneous as job applicants are already endowed with a differing set of skills.

2.3. Job recommending systems

Recommender systems provide personalized recommendations in order to alleviate information overflow Lu et al. [2015]. Many recomnender systems have been developed recently (e.g. Son and Kim 2018, Sun and Lee 2017, Liu et al. 2021, Asani et al. 2021, Viniski et al. 2021) and job market has not been an exception, as another stream of the literature looks for an optimal set-up of a job recommending systems (for an overview of the first studies see Al-Otaibi and Ykhlef [2012]). Job recommending systems are information systems supporting the recruitment process used by online recruitment platforms facilitated either by the recruiting company or an external provider. The design of a recommending system assumes the availability of information about the job opening (OJVs) combined with the information about job applicants (CVs). OJV data, analysed here, present one side of such database, with data on job openings but with only limited information on the side of job applicants (their views, reactions to OJVs). Reusens et al. [2018] evaluate a job recommending system algorithm designed to support career counselling and individual career decisions in the Flemish public employment service. Similarly, in Reusens et al. [2017] authors have studied what variables are relevant for job seekers' vacancy interest. Although we do not explicitly evaluate a job recommending system, our study is also closely related to this stream. Our data does not include information on individuals searching for a job. Therefore, instead of evaluating a two-directional recommendation algorithm, we consider OJV attractiveness over an unspecified (unrestricted/open) population of individuals searching for a job.

Gutiérrez et al. [2019] and Charleer et al. [2019] design a user-centred, interactive dashboards to explore job recommendations, where users can study what kind of changes in their skill-set might lead to different job recommendations. Finally, Frid-Nielsen [2019] relies on the popular gradient-boosting decision tree method to create a job recommending system for job applicants.

2.4. Contribution

While our study fits into the emerging literature based on utilizing OJV data, as far as we are aware, we are the first to study the attractiveness of OJV. Specifically, we measure attractiveness via three indicators, number of views that an OJV attracts, number of (unique) reactions (filling out the application form) that an OJV initiates on the side of a job-seeker, and the conversion rate. Our research thus contributes to the existing OJV literature and is of interest to various stakeholders of the labor market, including e-commerce operators of web-based job market platforms. *First*, employers are interested in designing OJV that will have the potential to attract highest possible attention (advertisement), reactions (more reactions increase the chance to select the right person for the job) and conversions (a higher conversion suggests successful targeting of the OJV). Our study provides strong evidence that non-linear models are much more effective in predicting views, reactions and conversions of OJV and that the most relevant variables are related to the job classification and job description characteristics. *Second*, OJV platform operators are interested in improving their services by providing tools that will help potential employers (job-seekers) to target their audience (potential employers). Outcomes from our study provide compelling evidence that such tools can be created using standard machine-learning algorithm (random forest). *Third*, policy makers are interested in reducing the miss-match between the supply and demand on the labor market; as suggested by Turrell et al. [2018a], reduction in the labor market miss-match might increase aggregate productivity. Our approach offers tools to connect job-searchers and job-offers more efficiency, thus leading to a reduction of labor market miss-match.

3. Data

We use data provided by a commercial e-recruitment company enjoying a dominant position on the market of a small European country - Slovakia. The company offers job advertising to employers recruiting for vacant positions. Employers pay for posting an OJV note, that is being published for 30 days. After 30 days, the OJV note needs a renewal which is linked to an additional payment. Employers are therefore motivated to attract as many relevant applicants over the 30 day period as possible. The OJV note can be withdrawn earlier if requested. Our data contains individual-level information on OJV that covers information on the type, location, occupation or requested skills coded via a predefined semi-opened list of options. Description of the offered job is available as well. Our data initially consists of 249812 cases covering a period from March 2018 until February 2019. Our sample is further reduced by filtering for jobs offered in a broader region of the capital city of Bratislava, where the job-seeker, applicant, had the option to fill-out an online application form directly on the website of the e-recruitment provider. The final sample consists of 32482 cases. From this sample, 80% cases are randomly selected to form a training sample and the remaining 20% is left for an out-of-sample testing.

3.1. Key performance measures: views, reactions and conversions

Following the distinction of Reusens et al. [2018], job-search is considered an active information retrieval: the user triggers the retrieval by consciously providing a search query. Our data includes two measures of OJV attractiveness collected in a two-stage process. First, some of the OJV is displayed/viewed as a result of the active information retrieval and a subsequent, an intentional click at one of the retrieved items. Let $i = 1, 2, ...$ denote individual OJV and V_i number of recorded OJV views. In the second stage, after the OJV is viewed, the user might decide to use the e-recruitment system to react to the OJV note by sending his CV. This interaction is called a 'reaction' and is denoted as R_i.

A high number of views can be perceived as an indicator of high visibility of the employer on the job market; ability to reach out to the large audience. Employers can also study this measure to evaluate their relevance on the job market. Each view (V_i) and reaction (R_i) presents an implicit feedback-based data collected under the information system that is running the job-search web portal. The number of reactions expects an even more active manifestation of

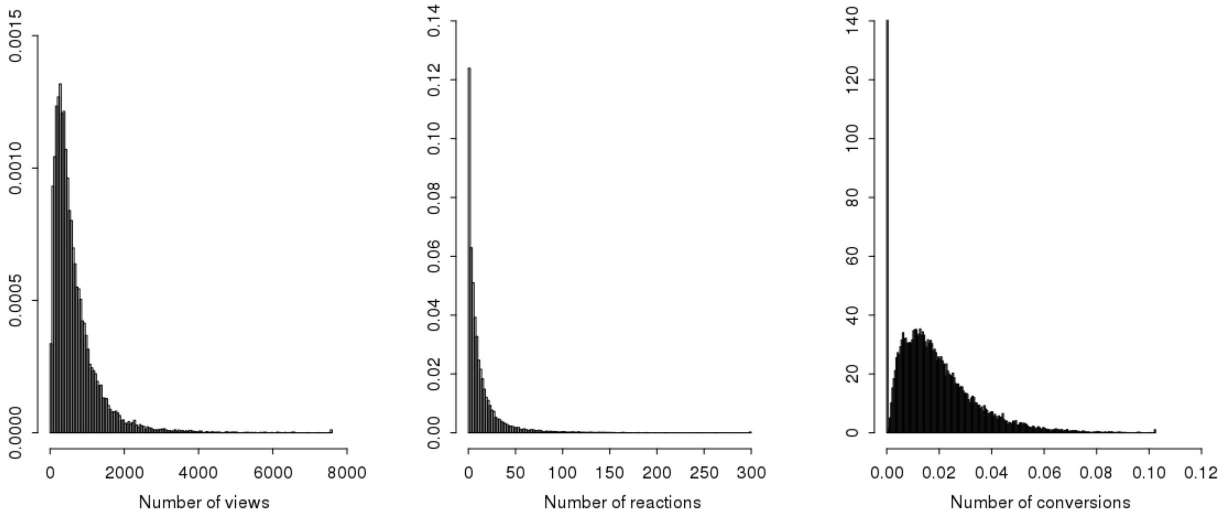

Figure 1: Distribution of views, reactions and conversions of online job vacancies

users' interest in a particular OJV. Views and reactions might vary strongly across different occupations, therefore out of these two measures of OJV attractiveness, we calculated the conversion, C_i:

$$C_i = \frac{R_i}{V_i} \in [0, 1] \tag{1}$$

Conversion scales reactions by views and should therefore be less sensitive to the presence of heterogeneity in attractiveness across occupations.

The following Figure 1 shows that all three measures are subject to right-skewness that is attributed to the presence of several outlying observations. Given the observed distribution, modeling $log(x + 1)$ transformation would be a possibility as well. However, we opt to model the raw data, as it is interpretable for both provider of the web-platform and employers. Moreover, predicting the log-transform requires an inverse conversions. Given Jensen's inequality, naive transformation of predicted values would lead to biased predictions, while more elaborate transformation (e.g. Taylor [2017]) are likely to introduce additional uncertainty.

In Table 1 we report key descriptive characteristics that show that on average OJV in our sample have attracted 655.73 views, with a considerable variation of 662.43 (standard deviation), with an upper quartile of 835.75, but the maximum at 12306. Reactions show a very similar distributional pattern. On average, 14.34 reactions are found for online job postings with an upper quartile at 17 and a maximum at 462. Views and reactions show a strong and statistically significant (with p-values $< 2.2 \times 10^{-16}$) dependence as suggested by Pearson's (0.75) as well as Spearman's (0.78) association coefficients. Finally, on average, 1 in 54 views lead to a reaction, which is captured by the average conversion at 0.0185, with the best case-scenario at 0.1335 or 2 in 15 views.

Table 1: Descriptive summary

Variable	Mean	Min.	Q1	Median	Q3	Max.	SD	Skew.	Kurt.
Views	655.73	0.00	257.00	470.00	835.75	12306.00	662.43	3.82	28.57
Reactions	14.34	0.00	3.00	7.00	17.00	462.00	23.25	5.25	47.00
Conversions	0.0185	0.00	0.0079	0.0156	0.0256	0.1335	0.0146	1.3371	2.7626

Note: Min. is the minimum, Q1 denotes a lower quartile, Q3 denotes an upper quartile, Max. is the maximum, SD is the standard deviation, Skew. and Kurt. denote skewness and kurtosis, respectively.

3.2. Overview of explanatory variables

For each online job vacancy, we have data for 883 potential explanatory variables. However, in many cases we have no or a very little variance and therefore after applying data cleaning procedures (described in the next Section), we are left with 172 (views) to 175 (reactions) features[2] These variables are stacked into 13 categories as shown in the following Table 2. A detailed list of these variables is found in Appendix. Here we present a short description of these variables:

- Benefits - benefits offered with the job (e.g. car, vouchers,..).
- Business sector - represents the field or industry in which the job is realized (e.g. banking, restaurants,...).
- Calendar effects - includes variables capturing specifics of the date when the OJV is posted (e.g. number of holidays in the next 30 days, monthly and weekly effects,...).
- Contract type - captures whether the position is a full-time job, part-time job, temporary agreement-based job, trade licence or internship.
- Education - includes different levels of required education (e.g. secondary with school-leaving examination, university education,...).
- Fresh graduate - only includes one dummy variable corresponding to 1 if the job is suitable for graduates.
- Language used in job posting - some postings are published in other than the Slovak language.
- Recruitment agency - includes only one dummy variable returning 1 if the job is posted by a recruitment agency.

[2] For each attractiveness measure, new training and testing samples were randomly generated.

- Reposting - includes only one dummy corresponding to 1 if the given OJV is a repost, i.e. it is open for longer than 30 days.
- Salary - includes only one dummy corresponding to 1 if the OJV includes an information about Salary.
- Skills - includes variables that correspond to different required skills (e.g. programming languages, driving licence,...)
- Text - includes variables that capture a simple morphological analysis of the description of the position (e.g. length of the job title in terms of letters, words, length of the description of the job,...).
- Job classification - corresponds to positions in the Unit group, a hierarchical level based on the ISCO classification of occupations (4-digit level).

Note that our data does not include information on the user performing the search. In explaining the attractiveness of the OJVs, we rely exclusively on the information present in the OJV note, including only information on the advertised job position and employer. Such data-based restriction might limit the predictive power of our models, but at the same time makes our results also relevant in the context of OJV data aggregators collecting information from multiple OJV advertisement providers (e.g. `https://www.burning-glass.com/`).

Table 2: Number of variables per categories

	Views	**Reactions**	**Conversion**
Category	No. of variables	No. of variables	No. of variables
Benefits	27	27	27
Business sector	23	23	23
Calendar effects	21	21	21
Contract type	4	4	4
Education	8	8	8
Fresh graduate	1	1	1
Job classification	47	49	48
Language used in job posting	2	2	2
Recruitment agency	1	1	1
Reposting	1	1	1
Salary	1	1	1
Skills	24	25	24
Text (job description) characteristics	12	12	12
Total	172	175	173

Note: Numbers of variables for each category are similar but differ, as for each attractiveness measure a new training and testing sample was generated.

3.3. Data cleaning procedures

Working with big data requires a careful handling of data. Using all data without filtering would lead to highly sensitive results or even to inability to estimate several models. We therefore applied following filters on the Training dataset:

- We removed all explanatory variables that had 0 variance. Most of our variables are dummy variables that indicate a certain skill, job classification, benefit, etc., as we are using a sub-set of the whole data, it might be that certain job characteristics do not appear in our training dataset.
- Given Spearman's correlation, $\rho^S_{j,k}$ $(j \neq k)$, calculated between all pairs of explanatory variables (j,k), we found those pairs j,k, where $|\rho^S_{j,k}| > 0.95$. We randomly selected one that was subsequently removed from our set of explanatory variables.
- We required dummy variables to have at least 0.5% of $1s$ or $0s$, whichever occurs less. Otherwise such variables were removed from our set of explanatory variables.
- Finally, we checked for the presence of exact multiple collinearity between all explanatory variables. If such variables are identified, they are removed from our set of explanatory variables, as they can be explained by a linear combination of other variables.

4. Methodology

In this section we first present models that we use to predict views, reactions and conversions. Next, we outline forecasting procedures and evaluation techniques.

4.1. Simple (un)conditional averages

Let $V_{i,s,n}$ denote i^{th} view belonging to the training set $i = 1, 2, ..., T$, s one of business sectors to which the OJV belongs, $s = 1, 2, ..., S$, where $T(s)$ is number of cases that belong to the given s^{th} business sector. One of the job classifications is denoted as $n = 1, 2, ..., N$ and as before, $T(n)$ is the number of cases that belong to the given n^{th} job classification. We use three benchmarks that are based on following averages taken over training samples:

$$\hat{V}^{b1}_{j,s,n} = T^{-1}\sum_{i=1}^{T} V_{i,s,n}; \quad \hat{V}^{b2}_{j,s,n} = T(s)^{-1}\sum_{i=1|s}^{T(s)} V_{i,s,n}; \quad \hat{V}^{b3}_{j,s,n} = T(n)^{-1}\sum_{i=1|n}^{T(n)} V_{i,s,n} \qquad (2)$$

The first, $\hat{V}^{b1}_{j,s,n}$ is a simple average over all observations in the training sample. The second, $\hat{V}^{b2}_{j,s,n}$ is an average over all cases in the training sample that also belong to the s^{th}

sector. Finally, given the results of Marinescu and Wolthoff [2020] who showed that variation across job classifications is the largest, we also use $\hat{V}^{b3}_{j,s,n}$ which is an average over all cases in the training sample that also belong to the j^{th} job classification. The three benchmarks lead to a simple, in case of $\hat{V}^{b1}_{j,s,n}$ uninformed, forecasts of views. Outperforming these forecasts allows us to access the true merit of more advanced techniques.

4.2. Linear models: OLS, LASSO, Ridge and Elastic Net

As an alternative to (un)conditional averages, we estimate and subsequently predict views, reactions and conversions using OLS, LASSO, Ridge and the Elastic Net models. The following optimization problem nests all four models Tibshirani [1996], Zou and Hastie [2005]:

$$\min_{\beta_0,\boldsymbol{\beta}} \frac{1}{2N}\sum_{i=1}^{N}(V_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\left[\frac{1-\alpha}{2}\sum_{j=1}^{r}\beta_j^2 + \alpha\sum_{j=1}^{r}|\beta_j|\right] \tag{3}$$

As before, V_i denotes views (replacing V_i with R_i or C_i leads to the models for reactions and conversions respectively) of i^{th} online job vacancy, $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ are $r \times 1$ column vectors of the standardized explanatory variables and coefficients, respectively and $\lambda, \alpha \geq 0$. If we let $\lambda = 0$, the model breaks down to the popular standard linear regression model, that might still be the first modeling choice before more complicated models are considered. Constraining $\lambda > 0$ adds weight to the penalty term, leading to LASSO, Ridge or Elastic Net models (Zou and Hastie 2005). Although the three models lead to biased coefficient estimates, the potential benefit is to achieve a lower out-of-sample forecast error. This is achieved by reducing the effect of irrelevant variables. Constraining $\alpha = 1$, the penalty term becomes $\lambda\sum_{j=1}^{r}|\beta_j|$, which is the LASSO model of Tibshirani [1996]. In this specification, variables that have a very little effect on the outcome V_i will receive zero weight, i.e. the coefficients will be equal to 0. At the same time, for highly correlated variables the LASSO tends to choose one (i.e. give a $\neq 0$ coefficient). Constraining $\alpha = 0$ leads to the ridge regression, where the penalty terms becomes $\lambda\frac{1}{2}\sum_{j=1}^{r}\beta_j^2$. In this specification, less relevant variables tend to receive a small but non-zero coefficient. Contrary to the LASSO, if two regressors are highly correlated, they both will receive a non-zero coefficient. Thus Ridge does not perform a variable-selection in the same sense as LASSO does. Finally, constraining $0 < \alpha < 1$ leads to the Elastic Net approach, which is a compromise between the two approaches. For LASSO and Ridge, an estimation of λ is carried out using the 10-fold cross-validation, where optimum λ^{opt} is the one that led to the lowest mean square error. A similar method is used for estimating α in case of the Elastic net, where we run two 10-fold cross validations. For each α parameter from $\alpha \in (0.1, 0.2, ..., 0.8, 0.9)$ an optimum λ^{opt}_{α} is found (inner 10-fold cross-validation) and corresponding mean square error is recorded. This is repeated for all values of α values and across all sub-samples of the 10-fold cross validation (outer cross-validation). An α_{opt} is selected where the mean square error in minimized. Corresponding OLS, LASSO, Ridge and Elastic Net forecasts are denoted as $\hat{V}_j^{OLS}$, $\hat{V}_j^{LASSO}$, $\hat{V}_j^{Ridge}$ and $\hat{V}_j^{EN}$ respectively.

4.3. Random forest

As a non-linear alternative, we rely on the random forest. Let $T(\boldsymbol{x})$ denote a decision tree with $\boldsymbol{x}$ representing a vector of predictors. A random forest is a collection of K such tree predictions Breiman [2010], specifically we use an average:

$$\hat{V}_j = K^{-1} \sum_{k=1}^{K} T_k(\boldsymbol{x}) \tag{4}$$

, where V_j is the predicted view of j^{th} online job vacancy. The specific algorithm is listed below and follows Hastie et al. [2009]:

- for $\{k = 1, 2, ..., K\}$
  - Select a bootstrap sample $\boldsymbol{Z}^k$ from the training data set.
  - Build a random forest to the sample $\boldsymbol{Z}^k$, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size is reached (which is set to 5):
    - Randomly choose m explanatory variables.
    - Select the best performing variable and split point - via mean square error loss function.
    - Split the node.
  - Output is a collection of trees: $\{T_k(\boldsymbol{x}) : k = 1, 2, ..., K\}$.

Several hyper-parameters need to be tuned. Specifically, the number of variables selected for a split at in each node $m \in \{20, 40, 60\}$, and number of trees $K \in \{1000, 3000, 5000\}$. We use 10-fold cross-validation over the full grid of possible hyper-parameter values. Optimum hyper-parameters correspond to those that minimized mean squared error. The corresponding forecasts are denoted as $\hat{V}_j^{RF}$.

4.4. Forecasting procedure and evaluation

We rely on a standard forecasting procedure, where 80% of all observations are left in the training sample for the purpose of estimating and tuning forecasting models. Remaining 20% observation belong to the testing sample, that are left to evaluate the accuracy of forecasts.

Let V_j denote views from the testing sample $j = 1, 2, ..., J$. For forecast evaluation purposes, we use two loss functions, absolute error:

$$AE_j = |V_j - \hat{V}_j| \tag{5}$$

and square error:

$$SE_j = \left(V_j - \hat{V}_j\right)^2 \tag{6}$$

To evaluate prediction accuracy of different models, we take the average across all predictions, which leads to the mean absolute error, $MAE = J^{-1} \sum_{j=1}^{J} AE_j$ and mean square error, $MSE = J^{-1} \sum_{j=1} SE_J$.

Differences between model accuracy might be negligible or just within a range of expected random fluctuations. We therefore test a hypothesis that forecast errors of one or more models are superior, i.e. lower, as opposed to those generated from remaining models. We use the model confidence set (MCS) approach of Hansen et al. [2011]. The MCS starts by assuming that forecasts from all models have same predictive accuracy. The set of these models is denoted as $\hat{Q}^0$ and consists of q models. The goal is to test, whether at a given confidence level $1-\alpha$, a subset of superior models exists, that we denote $\hat{Q}^*_{1-\alpha}$ with $q^* \leq q$ models. For each case in the testing sample, we compute loss differential. For simplicity, we assume square error loss (same approach applies for absolute error losses):

$$d_{uv,j} = SE_{u,j} - SE_{v,j} \tag{7}$$

where $u, v = 1, 2, ..., q$. Now calculate the average loss differential between model u and remaining models for j^{th} case:

$$d_{u\cdot,j} = \frac{1}{q-1} \sum_{v \in Q} d_{uv,j} \tag{8}$$

Now if we denote the expected value of these loss differentials as $c_{u,\cdot} = E(d_{u,\cdot})$, the hypothesis of interest is (Bernardi and Catania 2018):

$$\begin{aligned} H_{0,Q} &: \ \forall u = 1, 2, ..., q \quad c_{u,\cdot} = 0 \\ H_{1,Q} &: \ \exists u \text{ such that} \quad c_{u,\cdot} \neq 0 \end{aligned} \tag{9}$$

We use the $T_{max,Q}$ statistics given as:

$$T_{max,Q} = \max_{u \in Q} \frac{\bar{d}_{u,\cdot}}{\sqrt{\hat{D}(\bar{d}_{u,\cdot})}} \tag{10}$$

where $\bar{d}_{u,\cdot} = (q-1)^{-1} \sum_{v \in Q} \bar{d}_{uv}$ and $\hat{D}(\bar{d}_{u,\cdot})$ is an estimate of the variance of $\bar{d}_{u,\cdot}$. The variance is estimated via bootstrapping from a sample of losses.

The testing described above is the first step in a sequential procedure, where in each step the worst model is eliminated until the null hypothesis cannot be rejected. The final set of models belongs to a set of of models with a superior predictive ability with a given confidence level $1 - \alpha$. This approach has the benefit that it accounts for multiple hypothesis testing, thus is not that much sensitive to data-snooping bias. The usual confidence levels found in the literature range from 75% to 95%, we consider $\alpha = 0.05$.

4.5. Testing importance of variable groups

It is likely that not all variables have the same predictive power. We decided to test the predictive ability of each of the 14 variable categories described in Section 3.2. Our strategy is based on the MCS test described above.

Let the average loss (say mean square error) of the best forecasting model be MSE_{All}. Next, we run the forecasting procedure again, but this time, we exclude variables from a given category. For example, if we exclude benefits from our set of explanatory variables, the resulting average loss is denoted as $MSE_{Benefits}$. The larger the decline in the forecasting accuracy, the more important the Benefits for predicting the given variable are (views, reactions, conversions). The importance of variables belonging to the group g is thus calculated as:

$$I_g = 100 \times \frac{MSE_g - MSE_{All}}{MSE_{All}} \tag{11}$$

The statistical significance is assessed via the MCS procedure, where forecasting accuracy is compared. However, the losses are multiplied by $\times -1$, because if an exclusion of a given group of variables leads to increase forecast errors, it suggests that the given variable group is actually relevant.

An alternative to our approach might be a permutation based variable importance test, the BORUTA, approach of Kursa et al. [2010] based on the ideas of Stoppiglia et al. [2003]. In such methods, the individual variable importance is determined by comparing the predictive accuracy achieved by using the given variable and its randomized version. Using the MCS procedure, we can compare multiple models at once, thus also addressing the data-snooping bias and therefore we prefer the use of the model confidence set of Hansen et al. [2011].

5. Results

5.1. Predictability of job's attractiveness

In Table 3, we report forecasting accuracy results for views, reactions and conversions across two loss functions (Panel A and Panel B) and across all forecasting models. For example, the value of 370.388 in the first row and column is the mean square error of a completely uninformed analyst who is predicting views of all new online job vacancies using historical average, calculated over the training sample. The value of 371.311 in the second row is slightly higher and corresponds to the square forecast error for predictions generated by historical averages that correspond to the views of online job vacancies from the same business sector. A † symbol is placed next to models that belong to the superior set of all models. For example, in the first column of Panel A, all eight models were jointly tested via the model confidence set of Hansen et al. [2011] for the presence of a superior set of models. Only one model, the random forest is selected.

Results in Table 3 show several outcomes that hold across all three attractiveness measures and loss functions. First, a job classification seems to be a much better predictor as a business sector. This is partly expected given the results of Marinescu and Wolthoff [2020], where the job title seems to discriminate well between job offers. Opposed to the historical average benchmark, recorded improvements via mean square error are 13.89% (10.23% via mean absolute error), 8.53% (7.22%) and 14.09% (8.66%) for views, reactions and conversions respectively. Second, exploiting other OJV characteristics leads to further improvements in predictive accuracy. Even linear models such as OLS, LASSO, Ridge and Elastic Net show average improvements of 23.18% (15.98%), 13.33% (10.58%) and 18.64% (11.31%) for views, reactions and conversions respectively. However, even though our specifications start with over 170 explanatory variables, the shrinkage methods do not seem to provide any material advantage. Instead, our third key observation shows that the random forest out-performs all models achieving the highest improvements of 49.28% (36.79%), 35.30% (30.27%) and 35.01% (22.98%) for views, reactions and conversions. These improvements are also statistically relevant as predictions from the random forest model belong to the super set of models as indicated by the test of Hansen et al. [2011].

These results suggest that exploiting characteristics of online job vacancies and machine-learning methods can be used to improve predictability of OJV's attractiveness. It therefore follows that the job-filling rate, a key parameter in labor market matching models, can be

Table 3: Forecast errors: Predicting online job vacancy's attractiveness

	Views		**Reactions**		**Conversions**	
Panel A: Mean square error						
Benchmarks						
Unconditional historical mean	370.388		499.769		0.214	
Business sector specific mean	371.311		501.511		0.214	
Job classification	318.946		457.153		0.184	
Competing models						
OLS	285.097		433.753		0.174	
LASSO	284.323		432.711		0.174	
Ridge	284.550		433.363		0.174	
Elastic Net	284.217		432.769		0.174	
Random forest	**187.868**	†	**323.338**	†	**0.139**	†
Panel B: Mean absolute error						
Benchmarks						
Unconditional historical mean	419.554		12.713		11.304	
Business sector specific mean	420.351		12.740		11.303	
Job classification	376.614		11.795		10.325	
Competing models						
OLS	353.555		11.408		10.026	
LASSO	351.938		11.341		10.023	
Ridge	352.705		11.376		10.027	
Elastic Net	351.821		11.345		10.025	
Random forest	**265.219**	†	**8.865**	†	**8.706**	†

Note: Mean square errors for Views are divided by 1000, while both mean square errors and mean absolute errors for Conversions are multiplied by 1000. Values with the † symbol denote models that belong to the superior set of models at the 95% confidence level.

'manipulated', preferably increased, if data and methods are properly utilized.

5.2. Attractiveness of online job vacations: What matters?

Results from the previous section do not shed any light upon which variables are actually useful. We re-run our forecasting exercise, but now instead of using all variables we always remove those that belong to one of our feature categories (see Table A1). In this section, all results are based on the random forest model that led to lowest forecast errors. The idea being that if adding a variable (or a group of variables) will improve forecast errors, the given variable is likely to be important for managing online job's vacancy attractiveness.

In Table 4, we report forecast improvements. Specifically, a positive number of 7.8% in the first row and columns means that including job benefits, to all other variables, resulted in improvement of forecasting accuracy by 7.8%. Negative values actually suggest deterioration of forecasting accuracy, i.e. given variables add more noise than signal. As before, we observe consistent results across different measures of OJV's attractiveness. For views, reactions and conversions, the most relevant variables are *text* characteristics. These characteristics include the title of the job offer - its length in terms of words, number of words and number of unique words, and also the field tasks - its number of letters and length of tasks words. In terms of mean squared error, improvements are 23.1% (13.7% for mean absolute error), 15.3% (10.6%) and 10.1% (6.2%) for views, reactions and conversions respectively. The second most relevant variable is the *job classification*, which corresponds to a set of dummies. As before, these improvements are statistically significant and correspond to 16.1% (10.6%) for views, 12.4% (10.9%) for reactions and 9.6% (5.6%) for conversions. Note that one cannot simply add forecasting improvements, i.e. inclusion of *text* and *job classification* characteristics does not lead to a 23.1% + 16.1% forecast improvement. The reason is that the two variables are likely correlated. The role of these two variables are not that much surprising, as both Turrell et al. [2018a] and Marinescu and Wolthoff [2020] already showed that the job classification discriminates between different job characteristics well.

Interestingly, *job benefits* also systematically improve forecasting accuracy. Improvements are particularly notable for views (7.8% for MSE and 4.4% for MAE). As opposed to text and job classification characteristics, here the employers have much better control over what benefits they will provide. Note that while *job benefits* do not belong to the set of superior models (as indicated by the MCS test of Hansen et al. [2011]), it does not mean that these improvements are not statistically significant from the rest. In fact, in the sequential MCS algorithm, *job benefits* were removed last (for views, conversions and reaction, as well as for mean square and absolute errors), while only *text* and *job classification* remaining and the

MCS test was unable to distinguish between these two. It follows that also *job benefits* can be perceived as a relevant driver of online job vacancy's attractiveness.

Table 4: Importance of variable groups for predicting online job vacancy's attractiveness

	Views		**Reactions**		**Conversions**	
Panel A Mean squared error % changes to forecasting accuracy against the all variable model						
Job benefits	7.8%		4.6%		5.4%	
Business sector	-1.4%		-1.9%		-0.7%	
Calendar effects	-2.0%		-2.0%		-1.1%	
Contract type	-0.5%		0.0%		-0.2%	
Required education	1.6%		2.0%		2.7%	
Graduate status	0.0%		0.6%		0.8%	
Job posting's language	0.9%		0.4%		0.3%	
Recruitment agency status	3.6%		1.2%		0.8%	
Reposting status	-0.6%		-0.1%		-0.5%	
salary information	0.0%		0.2%		0.2%	
Skills	-0.8%		-1.2%		-1.0%	
Text (job description) characteristics	**23.1%**	†	**15.3%**	†	**10.1%**	†
Job classification	**16.1%**	†	**12.4%**	†	**9.6%**	†
Panel B Mean absolute error % changes to forecasting accuracy against the all variable model						
Job benefits	4.4%		4.1%		2.8%	
Business sector	-1.7%		-1.5%		-0.5%	
Calendar effects	-1.5%		-1.9%		-1.1%	
Contract type	-0.6%		-0.5%		-0.1%	
Required education	1.0%		2.4%		1.1%	
Graduate status	0.1%		0.2%		0.5%	
Job posting's language	0.3%		0.1%		0.2%	
Recruitment agency status	2.5%		0.9%		0.3%	
Reposting status	-0.5%		-0.2%		-0.3%	
salary information	-0.1%		0.1%		0.2%	
Skills	-0.9%		-0.6%		-0.6%	
Text (job description) characteristics	**13.7%**	†	**10.6%**	†	**6.2%**	†
Job classification	**10.6%**	†	**10.9%**	†	**5.6%**	†

Note: Values with † denote models that belong to the superior set of models.

6. Conclusion and agenda for future research

6.1. Concluding remarks

In this study, we had two research questions. First, can we predict online job vacancy attractiveness? Second, what variables are useful in predicting attractiveness of online job vacancies?

In order to answer the first question, we have defined attractiveness via three measures. A number of (unique) views of the posted job vacancy, number of reactions, i.e. filling out the application forms and the ratio of reactions to views. Next, we randomly stratified our sample into a training (80% of observations) and testing sample. Using up to 175 explanatory variables (from an initial sample of 883) grouped into 13 categories, we used four linear models (OLS, LASSO, Ridge and Elastic Net) and one non-linear (random forest) model that were tuned on the training sample. Finally, by the means of mean square and absolute errors and via the model confidence set test of Hansen et al. [2011], we compared which model performs the best in the training sample. We found a very consistent picture, the random forest always beats other models irrespective of the attractiveness measure employed or the preference for a loss function. Moreover, we presented a clear evidence that compared to an uninformed prediction via historical average, modeling attractiveness of online job vacancies by the means of random forest reduces squared forecast errors by almost 50% for views, 35% for reactions and still 35% for conversions. It therefore appears that the modelling is worth the effort.

In order to answer the second question, we have performed a variable selection analysis, where instead of removing one variable at a time, we removed a group of variables. For example, we removed all variables related to the benefits offered in the job description. A subsequent decrease (statistically validated via the MCS test) in the forecasting accuracy is a strong indication that variables from that group are relevant in predicting online job vacancy's attractiveness. We found that the *job classification*, *text* characteristics of job's description and also that *job benefits* are shaping online job's attractiveness. For example, excluding *text* characteristics variables from the random forest model led to a reduction of forecasting accuracy by 23.1%.

Our results point to three contributions that can be summarized as follows:

- Our approach can be used by employers to optimize their online job vacancy posting in order to increase attractiveness and thus potentially improving job-filling rates.
- Our research also helps providers of online job vacancy posts to improve the design of job searching interface. Variables that are relevant should be part of the searching interface.
- Finally, our results also have a potential implication from a broader macroeconomic sense, as reduction of the miss-match on the labor market can improve aggregate productivity.

6.2. Agenda for future research and research limitations

In this study, we do not use data from job-searching population. Understating their characteristics might improve attractiveness of online job vacancies. For example, we might be interested in characteristics that increase the probability of a job-searcher to apply for a position. Discrepancies between his characteristics and job's characteristics might be an important driver.

Our discussions with the operator of the online job vacancy portal revealed two interesting points. First, operators might be interested not only in the fact that a specific group of variables are relevant, but also in the direction of the effect. Although our model allows to extract such specific information, it is not necessary to do so. The idea is being that as employers are filling out job characteristics, our model will indicate predicted views, reactions and conversions. For example, offering/mentioning specific job benefits might increase or in some cases even decrease attractiveness of the job offer. The user will observe changes in predicted views, reactions and conversions as he types/selects job's characteristics. Second, operators are also interested in 'lower-bounds' of attractiveness. Currently, we predicted expected values and they were evaluated via symmetric loss functions. An alternative might be to predict specific quantiles. For example, predicting $5\%^{th}$ and $95\%^{th}$ percentiles might be more indicative for the user and platform provider.

Finally, while the random forest worked well, there are other methods that might even further improve our ability to predict online job attractiveness. Specifically, the extreme gradient boosting forest, deep neural networks and support vector machines. However, tuning of these methods is extremely computationally intensive[3]. Future works might explore, whether more complex methods will be worth the extra effort.

[3] Particularly for larger samples. In our case, with a training sample of 32482 cases and $172 - 175$ features/explanatory variables, we were unable to tune deep neural networks or support vector machines in a reasonable time (couple of days) - using a the following system: AMD Ryzen Threadripper 2970WX 24 core, 64GB RAM on a 64-bit Windows OS.

Appendix

Table A1: List of variables

Category	Variable	Notes
Benefit	Transportation	
Benefit	Accommodation	
Benefit	Team	
Benefit	Premium pay	
Benefit	Allowances	
Benefit	Multisport	
Benefit	Meal voucher	
Benefit	Sick day	
Benefit	Home office	
Benefit	Flexible time	
Benefit	Fourteenth salary	
Benefit	Thirteenth salary	
Benefit	Language course	
Benefit	Mobile phone	
Benefit	Car	
Benefit	Laptop	
Benefit	Vacation	
Benefit	Young team	
Benefit	Career	
Benefit	Stability	
Benefit	International company	
Benefit	Team building	
Benefit	Supplementary pension saving	
Benefit	Financial benefits	
Benefit	Self-development	
Benefit	Social benefits	
Benefit	Sport	
Business sector	Administration	
Business sector	Car industry	
Business sector	Banking	
Business sector	Tourism, gastronomy, hotel business	
Business sector	Transport, haulage, logistics	

Table A1: List of variables

Category	Variable	Notes
Business sector	Economy, finance, accountancy	
Business sector	Electrical & power engineering	
Business sector	Information Technology	
Business sector	Management	
Business sector	Marketing, advertising, PR	
Business sector	Commerce	
Business sector	Insurance	
Business sector	General labor	
Business sector	Service industries	
Business sector	Construction & real estate	
Business sector	Mechanical engineering	
Business sector	Top management	
Business sector	Production	
Business sector	Medicine & social care	
Business sector	Customer support	
Business sector	HR	
Business sector	Education, science & research	
Calendar effects	No holidays in the next 30 days	
Calendar effects	1 holiday in the next 30 days	
Calendar effects	2 holidays in the next 30 days	
Calendar effects	3 holidays in the next 30 days	
Calendar effects	April	
Calendar effects	August	
Calendar effects	December	
Calendar effects	February	
Calendar effects	January	
Calendar effects	July	
Calendar effects	Jun	
Calendar effects	March	
Calendar effects	May	
Calendar effects	November	
Calendar effects	October	
Calendar effects	Friday	

Table A1: List of variables

Category	Variable	Notes
Calendar effects	Monday	
Calendar effects	Saturday	
Calendar effects	Sunday	
Calendar effects	Thursday	
Calendar effects	Tuesday	
Contract type	Agreement-based (temporary jobs)	
Contract type	Full-time job	
Contract type	Part-time job	
Contract type	Internship	
Education	Secondary without school-leaving examination	
Education	Follow-up/Higher Professional Education	
Education	Secondary with school-leaving examination	
Education	Secondary school student	
Education	University student	
Education	Bachelor's degree	
Education	Master's degree	
Education	Primary education	
Fresh graduate	Fresh graduate	
Job classification	Senior Government Officials	
Job classification	Policy & Planning Managers	
Job classification	Sales & Marketing Managers	
Job classification	Health Services Managers	
Job classification	Civil Engineers	
Job classification	Electrical Engineers	
Job classification	Pharmacists	
Job classification	Early Childhood Educators	
Job classification	Accountants	
Job classification	Financial Analysts	
Job classification	Management & Organization Analysts	
Job classification	Personnel & Careers Professionals	

Table A1: List of variables

Category	Variable	Notes
Job classification	Advertising & Marketing Professionals	
Job classification	Systems Analysts	
Job classification	Software Developers	
Job classification	Applications Programmers	
Job classification	Software, Applications Developers & Analysts	
Job classification	Lawyers	
Job classification	Civil Engineering Technicians	
Job classification	Electrical Engineering Technicians	
Job classification	Physical & Engineering Science Technicians	
Job classification	Dental Assistants & Therapists	
Job classification	Accounting Associate Professionals	
Job classification	Commercial Sales Representatives	
Job classification	Buyers	
Job classification	Chefs	
Job classification	Hotel Receptionists	
Job classification	Stock Clerks	
Job classification	Waiters	
Job classification	Building Caretakers	
Job classification	Shop Supervisors	
Job classification	Shop Sales Assistants	
Job classification	Contact Centre Salespersons	
Job classification	Service Station Attendants	
Job classification	Prison Guards	
Job classification	Motor Vehicle Mechanics & Repairers	
Job classification	Electrical Mechanics & Fitters	
Job classification	Bakers Pastry cooks & Confectionery Makers	
Job classification	Product Graders & Testers, excl. Foods&Beverages	
Job classification	Assemblers	
Job classification	Heavy Truck & Lorry Drivers	

Table A1: List of variables

Category	Variable	Notes
Job classification	Lifting Truck Operators	
Job classification	Cleaners, Helpers in Offices, Hotels, etc.	
Job classification	Mining & Quarrying Labourers	
Job classification	Kitchen Helpers	
Job classification	Process Control Technicians	Only reactions
Job classification	Shelf Fillers	Only reactions, conversions
Job classification	General Office Clerks	
Job classification	Cashiers & Ticket Clerks	
Language used in job posting	English	
Language used in job posting	Slovak	
Recruitment agency	Recruitment agency	
Reposting	Reposting	
Salary	Salary dummy	1 if salary is posted, 0 otherwise
Skills	English A1	
Skills	English A2	
Skills	English B1	
Skills	English B2	
Skills	English C1	
Skills	Invoicing 1	
Skills	Invoicing 2	Only reactions
Skills	Business correspondence 2	
Skills	Microsoft Excel 1	
Skills	Microsoft Excel 2	
Skills	Microsoft Outlook 1	
Skills	Microsoft Outlook 2	
Skills	Microsoft PowerPoint 2	
Skills	Microsoft Windows 2	
Skills	Microsoft Word 1	
Skills	Microsoft Word 2	
Skills	German A2	
Skills	German B1	
Skills	German B2	
Skills	German C1	

Table A1: List of variables

Category	**Variable**	**Notes**
Skills	Treasury 1	
Skills	Treasury 2	
Skills	Stock control 1	
Skills	Slovak C1	
Skills	Slovak C2	
Text	No. of letters in the field describing the tasks	
Text	Task words length	
Text	Title no. of letters	
Text	Length of words in the Title	
Text	1 word in the Title	
Text	2-3 words in the Title	
Text	4-6 words in the Title	
Text	7-10 words in the Title	
Text	11-15 words in the Title	
Text	More than 15 words in the Title	
Text	7-10 unique words in the Title	
Text	11-15 unique words in the Title	

References

Al-Otaibi, S. T. and M. Ykhlef (2012). A survey of job recommender systems. International Journal of Physical Sciences 7(29), 5127–5142.

Asani, E., H. Vahdat-Nejad, and J. Sadri (2021). Restaurant recommender system based on sentiment analysis. Machine Learning with Applications 6, 100114.

Bernardi, M. and L. Catania (2018). The model confidence set package for r. International Journal of Computational Economics and Econometrics 8(2), 144–158.

Boselli, R., M. Cesarini, S. Marrara, F. Mercorio, M. Mezzanzanica, G. Pasi, and M. Viviani (2018). Wolmis: a labor market intelligence system for classifying web job vacancies. Journal of Intelligent Information Systems 51(3), 477–502.

Breiman, L. (2010). Random forests. machine learning.

Cedefop (2019). The online job vacancy market in the EU Driving forces and emerging trends. Cedefop research paper; No 72..

Charleer, S., F. Gutiérrez, and K. Verbert (2019). Supporting job mediator and job seeker through an actionable dashboard. In Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19, New York, NY, USA, pp. 121–131. Association for Computing Machinery.

Colace, F., M. De Santo, M. Lombardi, F. Mercorio, M. Mezzanzanica, and F. Pascale (2019). Towards labour market intelligence through topic modelling. In Proceedings of the 52nd Hawaii International Conference on System Sciences.

Colombo, E., F. Mercorio, and M. Mezzanzanica (2019). Ai meets labor market: Exploring the link between automation and skills. Information Economics and Policy 47, 27–37.

Davis, S. J., R. J. Faberman, and J. C. Haltiwanger (2013). The establishment-level behavior of vacancies and hiring. The Quarterly Journal of Economics 128(2), 581–622.

De Pedraza, P., S. Visintin, K. Tijdens, and G. Kismihók (2019). Survey vs Scraped Data: Comparing Time Series Properties of Web and Survey Vacancy Data. IZA Journal of Labor Economics 8(1), 1–23.

Deming, D. and L. B. Kahn (2018, 1). Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals. Journal of Labor Economics 36(S1), S337–S369.

Fabo, B., M. Beblavý, and K. Lenaerts (2017, 8). The importance of foreign language skills in the labour markets of Central and Eastern Europe: assessment based on data from online job portals. Empirica 44(3), 487–508.

Frid-Nielsen, S. S. (2019). Find my next job: Labor market recommendations using administrative big data. In Proceedings of the 13th ACM Conference on Recommender Systems, RecSys '19, New York, NY, USA, pp. 408–412. Association for Computing Machinery.

Grinis, I. (2019, jun). The STEM requirements of "Non-STEM" jobs: Evidence from UK online vacancy postings. Economics of Education Review 70, 144–158.

Gutiérrez, F., S. Charleer, R. De Croon, N. N. Htun, G. Goetschalckx, and K. Verbert (2019). Explaining and exploring job recommendations: A user-driven approach for interacting with knowledge-based job recommender systems. In Proceedings of the 13th ACM Conference on Recommender Systems, RecSys '19, New York, NY, USA, pp. 60–68. Association for Computing Machinery.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. Econometrica 79(2), 453–497.

Hastie, T., R. Tibshirani, and J. Friedman (2009). Random forests. In The elements of statistical learning, pp. 587–604. Springer.

Kureková, L. M., M. Beblavý, and A. Thum-Thysen (2015). Using online vacancies and web surveys to analyse the labour market: a methodological inquiry. IZA Journal of Labor Economics 4(1).

Kursa, M. B., W. R. Rudnicki, et al. (2010). Feature selection with the boruta package. J Stat Softw 36(11), 1–13.

Liu, H., Z. Sun, X. Qu, and F. Yuan (2021). Top-aware recommender distillation with deep reinforcement learning. Information Sciences 576, 642–657.

Lovaglio, P. G., M. Cesarini, F. Mercorio, and M. Mezzanzanica (2018). Skills in demand for ict and statistical occupations: Evidence from web-based job vacancies. Statistical Analysis and Data Mining: The ASA Data Science Journal 11(2), 78–91.

Lovaglio, P. G., M. Mezzanzanica, and E. Colombo (2020). Comparing time series characteristics of official and web job vacancy data. Quality and Quantity 54(1), 85–98.

Lu, J., D. Wu, M. Mao, W. Wang, and G. Zhang (2015). Recommender system application developments: a survey. Decision Support Systems 74, 12–32.

Marinescu, I. and R. Wolthoff (2020). Opening the black box of the matching function: The power of words. Journal of Labor Economics 38(2), 535–568.

Nicole, G., B. Lochner, L. Pohlan, and G. J. V. D. Berg (2020). Does Online Search Improve the Match Quality of New Hires ? .

Pissarides, C. A. (2000). Equilibrium unemployment theory, second edition. MIT press.

Reusens, M., W. Lemahieu, B. Baesens, and L. Sels (2017). A note on explicit versus implicit information for job recommendation. Decision Support Systems 98, 26–35.

Reusens, M., W. Lemahieu, B. Baesens, and L. Sels (2018, jul). Evaluating recommendation and search in the labor market. Knowledge-Based Systems 152, 62–69.

Rogerson, R., R. Shimer, and R. Wright (2005). Search-theoretic models of the labor market: A survey. Journal of economic literature 43(4), 959–988.

Son, J. and S. B. Kim (2018). Academic paper recommender system using multilevel simultaneous citation networks. Decision Support Systems 105, 24–33.

Stephany, F. (2020, 10). Does it Pay off to learn a new skill? Revealing the economic benefits of cross-skilling.

Stoppiglia, H., G. Dreyfus, R. Dubois, and Y. Oussar (2003). Ranking a random feature for variable and feature selection. The Journal of Machine Learning Research 3, 1399–1414.

Sun, C.-Y. and A. J. Lee (2017). Tour recommendations by mining photo sharing social media. Decision Support Systems 101, 28–39.

Taylor, N. (2017). Realised variance forecasting under box-cox transformations. International Journal of Forecasting 33(4), 770–785.

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

Turrell, A., B. Speigner, J. Djumalieva, D. Copple, and J. Thurgood (2018a). Staff Working Paper No. 742 Using online job vacancies to understand the UK labour market from the bottom-up. Technical report.

Turrell, A., B. Speigner, J. Djumalieva, D. Copple, and J. Thurgood (2018b). Using job vacancies to understand the effects of labour market mismatch on UK output and productivity. Technical report, Bank of England - Staff Working Paper No. 737.

Viniski, A. D., J. P. Barddal, A. de Souza Britto Jr, F. Enembreck, and H. V. A. de Campos (2021). A case study of batch and incremental recommender systems in supermarket data under concept drifts and cold start. Expert Systems with Applications 176, 114890.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) 67(2), 301–320.