Estimations of Professional Experience with Panel Data to Improve Salary Predictions

Frank Eichinger[1], Jannik Kiesel[2], Matthias Dorner[1], and Stefan Arnold[2]

[1] DATEV eG, Nuremberg, Germany
frank.eichinger@datev.de, matthias.dorner@datev.de
[2] Friedrich-Alexander-Universität Erlangen-Nürnberg, Nuremberg, Germany
jannik.kiesel@fau.de, stefan.st.arnold@fau.de

Abstract. Predicting salaries is crucial in business. While prediction models can be trained on large and real salary datasets, they typically lack information regarding professional experience, an essential factor for salary. We investigate various regression techniques for the estimation of professional experience based on data from the Socio-Economic Panel (SOEP) to augment data sets. We further show how to integrate such models into applications and evaluate the usefulness for salary prediction on a large real payroll dataset.

Keywords: salary prediction · regression · socio-economic panel

1 Introduction

Salary predictions are important for employers and (prospective) employees alike. When it comes to salary negotiations, both sides need to know the market value of an employee. Many tools on the Internet offer free salary predictions. However, many of them build on data obtained by questioning their users. Such data is notoriously poor in quality, as self-reports are often biased or wrong by purpose. Other approaches are built on objective data from official notifications to state authorities or on data from payroll software. Examples of such tools include the "Gehaltsvergleich BETA" [5] of the German Federal Office of Statistics which makes salary predictions with a linear-regression approach [6] or the application "Personal-Benchmark online" [3] (see Figure 1) from the German company DATEV eG making predictions based on a neural network [4][1].

While salary predictions based on large amounts of real salary data certainly result in better predictions than other approaches based on less data of questionable quality, such datasets rarely come with all the information needed. While the profession and regional information is typically available as well as the age and education of an employee, the professional experience is frequently missing as in the two applications mentioned (but can be obtained by questioning users directly). However, professional experience is widely considered being a crucial factor for productivity, which – besides further factors – determines salaries [10].

[1] [4] builds on random forests, the current application is based on a neural network.

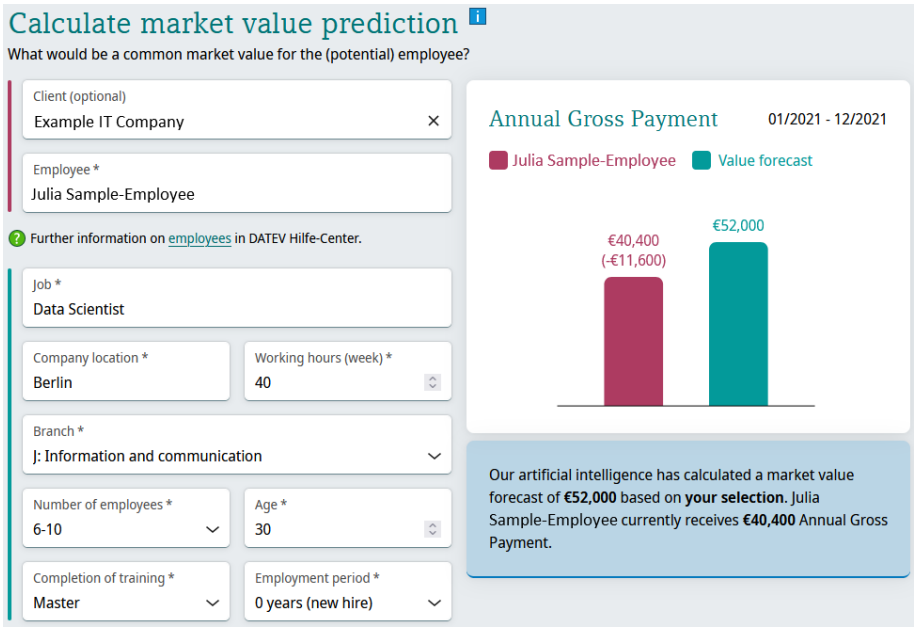

Fig. 1. Screenshot of "Personal-Benchmark online" from DATEV eG [3].

Further, the customers of DATEV eG request possibilities to enter this information. To improve predictions, it is desirable to augment and additionally consider the professional experience in datasets where it is missing.

Augmentation of information regarding professional experience to a dataset is not trivial, as this information is typically not available in large salary datasets, and crafting basic deterministic rules like "age minus default education time for reported degree minus 6" [10] would be too simple. Further, we are not aware of any research investigating the relationship between professional experience and other factors available in payroll data.

We present our approach for estimating professional experience based on external panel data. We use the dataset of the German Socio-Economic Panel (SOEP) [7] where a representative sample of employees has been questioned systematically. This dataset includes, among others, the age, the education and the professional biography. It is one of Europe's most important research datasets in the social and economic sciences. Our work contributes as follows:

1. We train and evaluate several machine-learning models for the estimation of professional experience on the SOEP panel dataset using variables available in payroll data and publish them to the public [8].
2. We present how a regression model for the augmentation of the professional experience can be integrated into an application for salary prediction.
3. We show in a large-scale evaluation with real payslip data from more than 4.3 million employees that salary prediction benefits from augmentation.

2 The Socio-Economic Panel and the Relevant Variables

The German Socio-Economic Panel (SOEP) [7] is an annual survey that has interviewed persons since 1984 systematically. It contains a representative sample of persons living in Germany and serves as a valuable resource for studying various socio-economic phenomena, particularly regarding social and labour-market policy, as well as income and career trajectories. We use the dataset published in 2022 [9] encompassing information from 13,616 individuals being employed at the time of interview (after omitting some data records in our preprocessing, e.g., where no education is captured or employees are older than the usual retirement age). Within this dataset, two specific subsets are of particular relevance:

- `bkpbrutto` – *Individual-specific data*: This subset contributes the essential variable *age* to our study, which we derive from `bkgeburt` (year of birth).
- `pgen` – *Occupation-specific data*: From this subset we derive the variable *professional_experience* in years, which serves as our target variable. We derive it from `pgexpft` (working experience in full-time employment) and `pgexppt` (working experience in part-time employment) with equal weight. Further, we determine *highest_education* based on `pgpbbil01` (vocational degree received), `pgpbbil02` (college degree), `pgpbbil03` (no vocational degree) and `pgisced11` (international standard classification of education). We adjust the values of *highest_education* to match the ones used in payroll software and official statistics in Germany as listed in Table 1.

Value	Description	Value	Description
1	no professional qualification	4	bachelor degree
2	vocational education	5	master degree
3	master craftsperson/tradesperson	6	doctoral degree

Table 1. The variable *highest_education* as used in official statistics in Germany. The description enumerates the typical representative but includes all equivalent degrees.

Overall, the SOEP dataset provides a robust foundation for investigating the interrelationships of the variables mentioned. By considering the *age* and the *highest_education*, we can estimate the variable *professional_experience*. Note that we explicitly do not investigate the variable *gender*, even if it would be available in the SOEP dataset and payroll data alike and would positively influence the quality of estimations of *professional_experience* as our preliminary experiments have shown. The reason is that this would contribute to the gender pay gap in the salary predictions we make in the next step.

3 Estimation of Professional Experience

In this section, we describe and evaluate different approaches to estimate the *professional_experience* using the variables *age* and *highest_education*. We encode

the *highest_education* using one-hot encoding and apply and evaluate several regression algorithms [1] using standard parameters: Linear regression, polynomial regression (degree 2–4), regression trees, random forests and neuronal networks with two hidden layers. Figure 2 displays two regression functions.

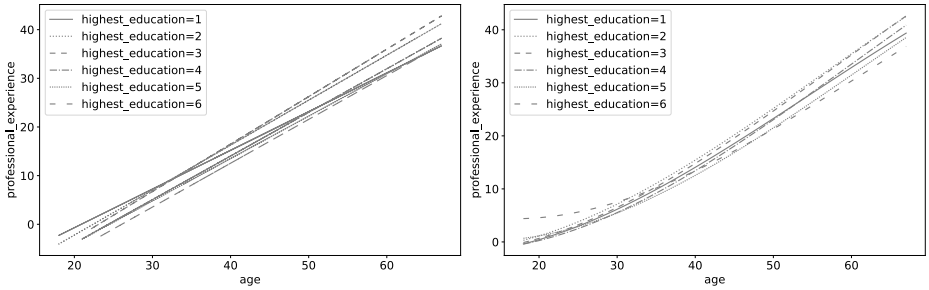

Fig. 2. A linear and a polynomial (degree 3) regression function (see Table 1).

We now present our experimental results with the different regression models in a 10-fold cross-validation setup measured by the standard mean absolute error (*MAE*). This measure is intuitive and makes sense from an application perspective.

Experiment	Technique	*MAE*
R1	Linear regression	3.90 years
R2	Polynomial regression (degree 2)	3.75 years
R3	Polynomial regression (degree 3)	3.74 years
R4	Polynomial regression (degree 4)	3.75 years
R5	Regression tree	3.78 years
R6	Random forest	3.78 years
R7	Neuronal network	3.72 years

Table 2. Experimental results in estimating the *professional_experience*.

Table 2 contains the results. The linear regression provides good results with an absolute error of 3.9 years. This seems to be acceptable, as biographies of employees are diverse, and estimations cannot be made without error (considering that they are built on two variables only). The polynomial regression techniques are more complex and lead to better results on all three metrics, with degree 3 as a winner. Surprisingly, the more advanced tree-based techniques perform a little worse. An explanation could be that these techniques have their strengths in settings with more variables. The neural network is the overall winner.

One advantage of linear and polynomial regression models – besides the prediction quality – is their simplicity [2]. Such a model is a simple mathematical

formula that can be easily integrated into any software application, regardless of the programming language. As the neural network performs only marginally better, we consider the polynomial regression with degree 3 as the best model for applications. We publish some of our models to the public [8].

4 Applicability for Salary Predictions

As motivated in the introduction, we aim at improving salary predictions such as [4] and [6] by augmenting salary datasets with an estimated variable describing the professional experience. This does not mean that we expect more accurate predictions (on a dataset where no real *professional_experience* is available), but that users can obtain predictions – as requested – based on variation introduced by the new variable *professional_experience*.

In this section, we prototypically integrate the published regression models [8] into the neural network for salary predictions of the application "Personal-Benchmark online" [3] as described in [4][1]. We perform augmentation of the new variable in a dataset of 4.3 million employees from the payroll software from DATEV eG from 2023 by applying the winning regression function of degree 3 from the previous section. To deal with missing *highest_education* information, we first impute this variable by using the most frequent value from employees having the same profession. We then train the existing neural network of the application with the additional variable.

We are particularly interested in the question if the *professional_experience* performs as well as a similar actual variable already present in the payroll data and used in the application, namely *employment_period* (see Figure 1). The *employment_period* is the time with the current employer in years, i.e., after changing the employer, it is 0. It has a Pearson correlation of 0,48 with the augmented *professional_experience* in our dataset. To ablate the benefits of the two variables, we conduct four experiments: with/without both variables and with both variables separately (see Table 3). We use a random 10% sample of the dataset for testing which we do not use for training. Table 3 contains the results with the mean absolute percentage error (*MAPE*) as evaluation metric as in [4].

Experiment	*employment_period*	*professional_experience*	*MAPE*
S1	–	–	15.17%
S2	–	✓	15.15%
S3	✓	–	14.94%
S4	✓	✓	14.90%

Table 3. Experimental results in salary prediction.

While Experiment S1 has a *MAPE* of 15.17%, S2 with the added estimated *professional_experience* performs almost equally. Adding the *employment_period*

in S3, however, improves the performance considerably. The reason is probably that a real variable describing the relationship of an employee with their employer better explains the salary than an estimated experience (which certainly is not correct in some cases). Adding *employment_period* additionally in S4 leads to an almost unchanged but slightly better result. We conclude that using both variables seems to be the best solution in terms of quality. The *professional_experience* however does not improve the results considerably in our test setup, where only estimated values are available, but allows the user to make better salary predictions based on this variable entered additionally as requested.

5 Conclusion

We have presented an approach for the estimation of the professional experience based on regression on data from the German Socio-Economic Panel (SOEP). We have shown how the variable can be integrated into software for salary prediction that is originally unavailable, and we have successfully evaluated the approach on a large real payroll dataset. Our plan is to integrate the solution into the application "Personal-Benchmark online". Our future research includes the investigation of regression models with more variables that are available in the SOEP and payroll data alike, such as school education, region and profession.

References

1. Aggarwal, C.C.: Data Mining: The Textbook. Springer (2015). https://doi.org/10.1007/978-3-319-14142-8
2. Cheng, X., Khomtchouk, B., Matloff, N., Mohanty, P.: Polynomial Regression As an Alternative to Neural Nets. CoRR in arXiv **abs/1806.06850** (2019). https://doi.org/10.48550/arXiv.1806.06850
3. DATEV eG: Personal-Benchmark online. `https://datev.de/web/de/mydatev/online-anwendungen/datev-personal-benchmark-online/`, accessed: 2023-07-10
4. Eichinger, F., Mayer, M.: Predicting Salaries with Random-Forest Regression. In: Alyoubi, B., N'Cir, C.B., Alharbi, I., Jarboui, A. (eds.) Machine Learning and Data Analytics for Solving Business Problems, chap. 1, pp. 1–21. Springer (2022). https://doi.org/10.1007/978-3-031-18483-3_1
5. German Federal Office of Statistics: Gehaltsvergleich BETA. `https://service.destatis.de/DE/gehaltsvergleich/`, accessed: 2023-07-10
6. German Federal Office of Statistics: Interaktiver Gehaltsvergleich. `https://www.destatis.de/DE/Service/Statistik-Visualisiert/Gehaltsvergleich/Methoden/Methodenbericht.pdf`, accessed: 2023-07-10
7. Goebel, J., Grabka, M.M., Liebig, S., Kroh, M., Richter, D., Schröder, C., Schupp, J.: The German Socio-Economic Panel (SOEP). Jahrbücher für Nationalökonomie und Statistik **239**(2), 345–360 (2018). https://doi.org/10.1515/jbnst-2018-0022
8. Kiesel, J.: Prediction Models for Professional Experience. `https://www.it-management.rw.fau.de/sgai/`, accessed: 2023-09-12
9. Liebig, S. et al.: Socio-Economic Panel, data from 1984–2020 (SOEP-Core, v37, Onsite Edition) (2022). https://doi.org/10.5684/SOEP.CORE.V37O
10. Mincer, J.: Schooling, Experience, and Earnings. National Bureau of Economic Research (1974)

Acknowledgement

This version of the contribution has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/978-3-031-47994-6_46. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms.