Prediction of Appointment Duration in Personal Services

Birger Lantow [0000-0003-0800-7939]

[1] University of Rostock, 18051 Rostock, Germany
Birger.lantow@uni-rostock.de

Abstract. The estimation of the duration of appointments is an important, often manual planning task. Efficiency and effectivity of work depend on the quality of these estimations. Idle times and appointments that do not reach their goals due to time limitations are the consequence of bad scheduling. For personal services, a lot of context information is required for estimations. Employees that schedule appointments may not have the possibility to process all context information or lack experience. An IT-based predictor for appointment durations could help here. This work investigates the feasibility of such a predictor by analyzing the application of first predictor models to a sample dataset from practice. First implications for this field are drawn from the result.

Keywords: Time Prediction, Personal Service, Data Science, Time Estimation.

1 Introduction

Time is a non-renewable resource, and it cannot be increased, preserved, or saved. In the context of scheduling appointments, the prediction of the duration of an appointment is an important task. If the appointment lasts shorter than expected, idle times may occur. If it lasts longer than expected, following appointments must be rescheduled, or the appointment will end without reaching its goal. Either way, coordination effort increases, and efficiency as well as effectivity of work decrease.

Usually, the duration of appointments is estimated based on personal experiences when is appointment is being planned. This is a complex task that involves, for example, the assessment of appointment goals, the background, and the number of involved participants or even the time of day. However, sometimes there is no experience or relevant context information available when planning an appointment.

What if data analysis can provide help here? In certain domains such as personal services provision, the duration of appointments, the topics, the participants, and other data are recorded for billing purposes. This work investigates the possibility of predicting the duration of appointments based on historical data from such records. This is a first step, evaluating the general feasibility. A neural network has been trained for the prediction of appointment durations and compared to an average based prediction as a baseline.

The remainder of this paper is constructed as follows. Section 2 discusses related work in the domain of machine learning based time prediction. From an industrial

perspective, the terms ‘time prediction’ or ‘time estimation’ are used here, referring to processing times or cycle times, for example. From the perspective of appointments in personal services, we use the term ‘appointment duration’. Section 3 provides a deeper insight into the domain of personal services and assumptions made for the estimation approach. In the following, Section 4 describes and evaluates the approach based on a prototypical application. The last section provides a summary and outlook.

2 Related Work

Forecasting activities on the basis of prediction of time play a very important role in several domains [1, 2]. Time prediction is one of the most challenging tasks in the context of forecasting and has received a lot of interest in research in recent years [2]. The accuracy of time forecasting can be fundamental to decisions in processes in today’s business world [2]. The subjectivity of assessments and the increasing importance of time forecasting are the reasons why tools for time forecasting are indispensable in today’s business [3].

We have studied various scientific publications related to predicting work time. These sources describe the prediction of working time in various subject areas such as agriculture and manufacturing.

Marco Fedrizzi et al. compared linear and non-linear approaches to predict the time it takes for an agricultural automatic production machine to complete a required field operation. An artificial neural network was used as a non-linear approach. The linear approach was represented by multiple linear regression. For the prediction, different forms of agricultural fields were considered, as well as different ways of processing these fields [4]. As a result, it was found that an artificial neural network gives much more accurate results than multiple linear regression in such subject areas. A further advantage of neural networks is the possibility to have categorical, continuous as well as discrete variables as inputs. This supports various possibilities to describe the context of recorded and estimated times.

Bozena Hola et al. presented a methodology for determining earthworks execution time and costs. Artificial neural networks were used to determine the productivity of selected sets of machines [5]. The investigations and analyses showed that the selected feed-forward multilayer error backpropagation neural network with a conjugate gradient algorithm is useful for predicting productivity.

Yu et al. [6] proposed clustered neural networks for the prediction of travel times. They achieved good prediction results based on a traffic simulation. Potentials for further development are seen in the improvement of neural network construction.

Phillip Backus et al. compared classification and regression tree, nearest neighbor, and artificial neural network algorithms for factory cycle-time predictions for lots. Their research showed that regression trees provide more accurate results in such subject areas. Regression trees are a flexible tool to define important variables and define similarity between lots [7]. Furthermore, the scientists concluded that variables need not be scaled to common units in such predictors. Predictions were improved by using

an unsupervised algorithm, such as clustering, to form similar groups of lots and then apply the tree algorithm to the clusters.

Yu and Cai [8] used support vector machines (SVM) for the prediction of working times in aircraft assembly. However, all input variables have been metric values.

In general, the time prediction approaches using neural networks show promising results and allow multiple kinds of variables as an input describing the context. Another promising approach are regression trees that have the same advantages as neural networks and showed a better accuracy compared to neural networks in the study of Backus et. al. However, regression trees tend to overfit, and small changes in the training data may result in big changes in the tree structure. For the first evaluation of machine learning based prediction of appointment durations, the neural network approach has been selected.

3 Problem Description

The common characteristic of Personal Services is that they are directed at persons instead of things. The recipient of the service is at the same time in the role of a co-producer. This situation as a key element in Personal Services reduces the predictability and increases the uncertainty in the corresponding work processes. Therefore, many Personal Service processes show characteristics of knowledge-intensive processes that come with a high variability and high autonomy of the actors. Fließ et al. [9] describe three phases or personal services: pre-service, service, and post-service. Personal Services are often integrated into a longer-term meta-process, which then also has long-term objectives. In this case, we speak of Long-Term Personal Services that involve multiple service encounters. Examples would be services in Physiotherapy, Psychotherapy, Family Care, or Coaching.

Since the nature of Personal Services is the interaction between people, human interaction and relationships play a strong role for service performance and thus should be considered when analyzing service processes. This long-term relationship has great importance, for example, in therapy and coaching settings. Looking at the long-term process, context changes are likely to influence the outcome as well. An example would be a changing life situation caused by a new occupation. Information with regard to context and relationship is commonly available in unstructured documentation, including, for example, diaries, written assessments, or communication content. [10]

From the perspective of the prediction of appointment durations, each service encounter is considered an appointment. The duration of a service encounter is influenced by the client, the representative of the service provider, and the current situation, including long-term and short-term aspects of the service process. All these influences provide input variables for the prediction of the appointment duration.

For simplification of the approach, we assume that each service encounter can be scheduled independently. Hence, the duration is not influenced by other appointments of the involved actors. Furthermore, we do not divide between planned (actual appointments) and unplanned service encounters. Not every phone call is planned in advance, and there might be crisis situations that require immediate action. Still, a prediction of

the duration of unplanned activities provides valuable information for rescheduling future appointments.

4 Prototypical Application

For the feasibility analysis of appointment duration prediction for personal services, we use a dataset of a small German family care company. The dataset is described in Section 4.1. Section 4.2 then portrays the data preparation and the implementation of the neural network for the prediction. This is followed by the evaluation of the approach in section 4.3.

4.1 Data Understanding

The company provides personal services like coaching and consulting that characterized a personal interaction between service provider and service consumer. Generally, service delivery is bound to appointments between both involved parties. The dataset contains the following features of each appointment (task):

`start` - Task execution start time
`stop` - Task execution stop time
`project_id` - Unique identifier of the client (service consumer)
`task_id` - Unique identifier of task type
`user_id` - Unique identifier of employee (service provider)

Task execution start and stop time are presented in timestamp format and consist of year, month, date, hour, and minutes (e.g. 01.03.2016 17:15). Each row of the dataset consists of information about only one appointment instance. An exemplary part of the dataset can be seen in Figure 1. Table 1 shows the task types that can be distinguished by the `task_id`. In total, the dataset contains 39,849 records that have been collected for billing purposes starting in 2016. The duration of an appointment can be calculated by the difference between stop and start timestamps. This results in multiples of 15 minutes.

	id	start	stop	project_id	task_id	user_id
0	15	01.03.2016 13:30	01.03.2016 17:00	62	3	73
1	17	04.02.2016 09:49	04.02.2016 10:49	41	3	82
2	18	02.03.2016 13:30	02.03.2016 16:30	5	3	73
3	19	03.03.2016 12:00	03.03.2016 16:30	158	7	87
4	20	04.03.2016 07:30	04.03.2016 09:15	91	3	78
5	24	09.03.2016 16:00	09.03.2016 17:00	41	3	82
6	27	01.03.2016 09:00	01.03.2016 09:45	44	3	75
7	28	01.03.2016 16:00	01.03.2016 17:15	153	4	75
8	30	02.03.2016 08:00	02.03.2016 10:00	111	5	75
9	31	03.03.2016 08:00	03.03.2016 10:30	5	3	73

Fig. 1. Example slice of the dataset

The original dataset also contained data like the names of additional actors that are involved and unstructured documentation of the appointment itself as well as achieved results or diagnostic information. These contents have been filtered out for data privacy reasons. Therefore, the data that has been available for this study provides only little context information. The influence of employee and client personality can be considered by a predictor because both are identified. The data contained 6088 unique combinations of `task_id`, `project_id`, and `user_id` – specific combinations of employee, client, and performed task type. This results in an average of 6.5 records per combination. The `task_id`s are not evenly present in the dataset. Figure 2 shows the distribution of `task_id`s in the data.

Table 1. Task types

id	German Name	English Translation
1	Telefongespräch	Phone Call
2	Therapie	Therapy
3	Begleitung	Supervision
4	Beratung	Counselling
5	Netzwerkarbeit	Networking
6	Hausbesuch	Visit at Home
7	Aktivität	General Activity
8	Dokumentation	Documentation
9	Hilfeplangespräch	Planning appointment
10	Krisenintervention	Crisis Intervention
11	Umgangsbegleitung	Contact Supervision
12	Fachaustausch	Professional Exchange
13	Praktische Unterstützung	Practical Support

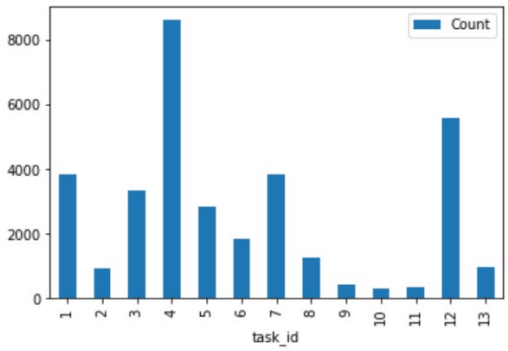

Fig. 2. Occurrences of the different task types in the data

The influence of time as an additional context element can be taken into account using the timestamps. Motivation and capacity of human actors change during a workday, during a working week and also depend on the season of the year. There is a non-linear effect. Hence, capacity does not continually decrease or increase during a day for example. However, several studies analyzing human performance show a periodicity in a 24 hours period (daily), a 7 days period (weekly), and a 365 days period (yearly).

Some examples of studies with regard to daily performance cycles can be found in [11]. Furthermore, business planning is generally based on such periods.

4.2 Data Preparation and Predictor Models

In order to predict the duration of an appointment, three predictor models have been investigated:

1. Average duration per task type as a baseline predictor
2. Average duration of a task type per employee and client. As a predictor that fits to the influence of different actors involved in the appointments
3. A neural network that has involved actors, task types, and time information as inputs as a machine learning based predictor

All data processing and prediction model implementation has been done using Python, pandas[1], and scikit-learn[2]. Using the toolset, some data preparation steps have been performed before the training and application of the predictor models. After removing records with null values, durations have been calculated as the difference between stop and start timestamps. Furthermore, combinations of `task_id`, `project_id`, and `user_id` that occurred only once in the data have been removed. Otherwise, the second average-based predictor would benefit in evaluation because of no prediction errors in these cases, although there is no real prediction having just one value. The boxplot in Figure 3 shows several outliers for the different task types. All records with values outside 1.5 IQR (InterQuartile Range) marked by the whiskers have been removed from the dataset. After these steps, 34091 records remained for training and evaluation.

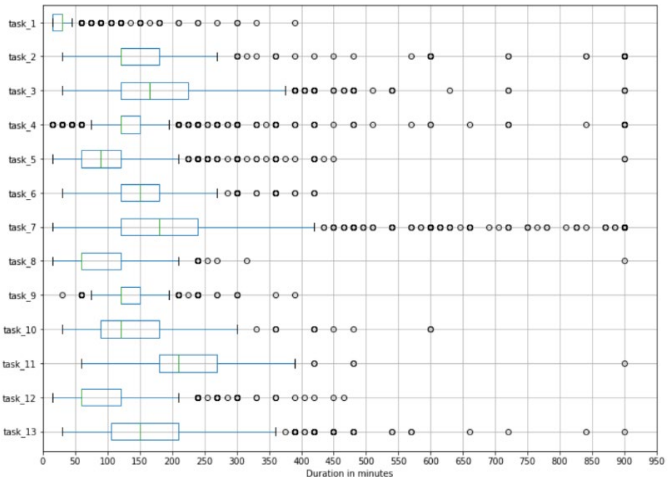

Fig. 3. Boxplot of durations per task type

[1] https://pandas.pydata.org/
[2] https://scikit-learn.org/

With regard to time information, the data is sparse (there are periods without records), small (2622 records per task type on average), and is assumingly overlaid by several other influence factors (e.g. different employees and clients). Therefore, simple periodicity features have been calculated in order to support prediction performance of the neural network approach. For the daily periodicity, a feature `daytime` (1 – morning, 2- noon, and 3-evening) has been introduced. For the weekly periodicity, a feature `weekday` symbolizing the seven weekdays has been added. And last, a feature `season` (1 – winter, 2 – spring, 3- summer, and 4 – autumn) is used for the yearly periodicity.

For training an evaluation, the data has been randomly split into a training dataset (70% of the records) and a test dataset (30% of records). Furthermore, the input data has been scaled using the StandardScaler provided by scikit-learn.

The MLPRegressor class has been used as the model for the neural network. In the network, there is one hidden layer, which consists of 600 nodes, and the learning rate is adaptive. That means that the learning rate is constant as long as training loss keeps decreasing. The adam solver is an adaptive learning rate optimization algorithm that has been designed specifically for training deep neural networks. Different configurations, e.g., with more hidden layers and a different number of neurons, but no significant change in the prediction performance (see Section 4.3) has been noticed.

4.3 Evaluation

The Mean Absolute Error (MAE) has been chosen for evaluation purposes. This helps to assess practical implications of the observed error by comparing it to the average duration of a certain task type. For the overall dataset, the neural network showed a MEA of 49.6 minutes, the task type-based prediction a MAE of 33.9 minutes and the prediction based on the average duration per task type, employee, and client a MAE of 21.7 minutes. Taking these values, the neural network is clearly outperformed by the average based prediction. The best result is realized when considering the specifics of involved actors in the average. However, as can be seen in Figure 2 and Figure 3, there is an imbalance in the occurrence of certain task types in the data and there are differences in the average duration of appointments per task type. Thus, prediction might be better for highly represented task types, and MEA should be seen in relation to the average duration of the specific task type. Table 2 and Figure 4 show the results.

It can be clearly seen that the average specific to employee, client, and task type (red bars) outperforms the other predictors for every task type. Even for the highly represented task types Phone Call, Counselling, General Activity, and Professional Exchange the neural network shows poor performance. For Phone Call the MAE is even bigger than the average duration. On the other hand, the best performing predictor shows, for example, an average error of roughly 6 minutes for Phone Calls that have an average duration of 26 to 27 minutes. However, it can also be seen that the performance varies depending on the task type.

Table 2. Mean duration and MEA per task type

Task Type	Mean Duration	MAE: NN Prediction	MAE: Average per Employee, Client, and task type	MAE : Average per task type
Phone Call	26.530.374	86.482.888	5.749.542	7.724.665
Therapy	134.340.836	27.142.918	12.947.640	27.902.109
Supervision	176.373.874	69.805.498	35.216.246	53.885.995
Counselling	132.006.961	23.856.116	16.122.649	21.404.178
Networking	94.817.480	39.222.999	23.615.666	37.394.670
Visit at Home	146.597.601	48.411.912	18.161.557	47.185.329
General Activity	172.894.530	74.609.332	36.718.127	64.393.122
Documentation	78.463.357	47.532.671	18.859.308	31.653.842
Planning Appointment	124.689.737	18.915.845	11.175.711	17.232.059
Crisis Intervention	127.630.662	44.503.165	24.279.832	42.999.795
Contact Supervision	220.221.239	112.059.537	29.683.531	61.106.195
Professional Exchange	79.299.158	44.722.698	23.734.403	30.379.892
Practical Support	142.351.408	49.419.122	29.772.634	50.026.225

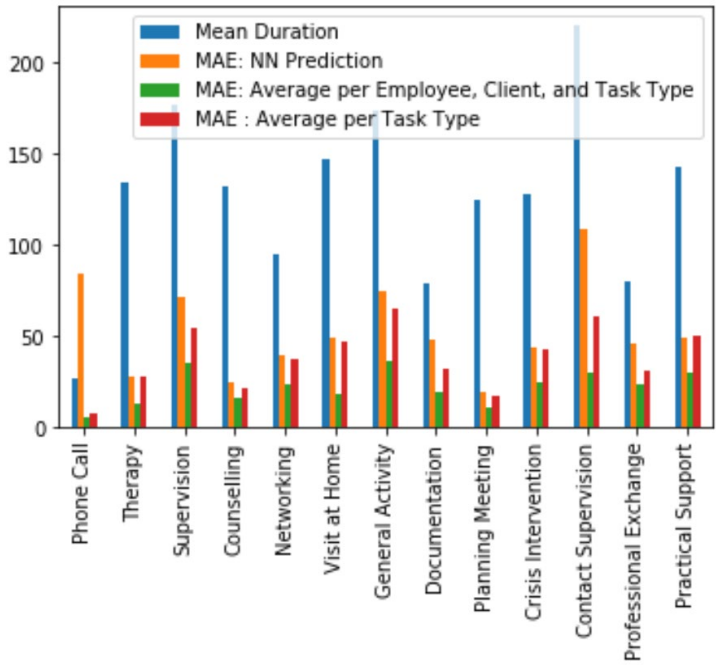

Fig. 4. Mean duration and MEA per task type

Overall, the neural network did not perform well on the dataset. A reason may be the high influence of the actors on the duration of the appointments. This high influence can be seen comparing the outcome of task type average predictor and the predictor that also considers the actors when calculating the average. Overall, the underperforming predictors underfit. Furthermore, the average duration of appointments per employee, client, and task type seems to be a good predictor depending on the task type.

5 Conclusion and Outlook

This study presents a first view on the topic of duration prediction for appointments. Some first assumptions can be made based on the findings. However, a validation using different data sources and a large dataset should be done in the future.

Time in terms of time of the day or day of the week does not seem to have a big influence on the duration of appointments in the domain. The good performance of the predictor based on the average duration per task type, employee, and client indicates a large personal influence on the appointment duration. Furthermore, this predictor could provide a simple implementation that fits practical requirements, at least for some task types. A validation is planned for the future.

The neural network approach clearly underperforms. Reasons may lie in the large influence of the involved actors. This results in categorical variables as inputs for the neural net that have a high number of possible values (ids of employees and clients). A point for future investigations is the big MAE that has been shown by the neural network for phone calls although there was a lot of training data available in comparison to other task types. It would be interesting to learn about the reasons.

The best performing average based predictor runs into problems for new combinations of employees, clients, and task types because there is no past data available that can be used for prediction. One scenario would be a new employee at the service provider. However, in this case, duration prediction might be very important, assuming a lack of experience for new employees. In the case of a new client, there is a lack of experience with the client. A straightforward solution would be the task type-specific average, but this comes with a significantly higher error for some task types (see Figure 4). Further approaches are the clustering of actors according to their characteristics, which may be done manually or automated data-driven or the inclusion of more complex context variables. Here the problems of weighting these factors and non-linear correlations pop up. This demands for a more complex prediction model. A neural network that combines these additional variables and the averages as inputs could be a solution. First tests with the current dataset showed a similar performance of neural network and the average based predictor.

Subsuming this discussion, additional data seems to be required in order to improve the prediction capabilities. First, with regard to the considered variables, this could be the status of the long-term personal services process, as suggested in [10], the number of participants in an appointment, or the personal characteristics of the actors. Second, with regard to the amount of data and generalizability, this could be a long-term data collection and a collection of data in different companies that is already underway.

At last, the assumption that appointments don't influence each other imposes a limitation that might reduce the potential of practical utility. Such problems might be revealed by a deeper analysis of the input data and its quality.

Acknowledgment

We thank the GeBEG Rostock GmbH for providing the anonymized dataset and for giving an insight into the company's work.

References

1. Aliev, Rafik, Bijan Fazlollahi, Rashad Aliev, and Babek Guirimov: Fuzzy time series prediction method based on fuzzy recurrent neural network. In International Conference on Neural Information Processing, pages 860–869. Springer, 2006.
2. Son, Nguyen Thanh, Nguyen Hoai Le, and Duong Tuan Anh: Time series pre- diction using pattern matching. In 2013 International Conference on Computing, Management and Telecommunications (ComManTel), pages 401–406. IEEE, 2013.
3. Khashei, Mehdi, Mehdi Bijari, and Gholam Ali Raissi Ardali: Improvement of auto-regressive integrated moving average models using fuzzy logic and artificial neural networks (anns). Neurocomputing, 72(4-6):956–967, 2009.
4. Fedrizzi, Marco, et al. "An artificial neural network model to predict the effective work time of different agricultural field shapes." Spanish journal of agricultural research 17.1 (2019): e0201.
5. Hola, Bozena, and Krzysztof Schabowicz. "Estimation of earthworks execution time cost by means of artificial neural networks." Automation in Construction 19.5 (2010): 570-579.
6. Yu, Jie, Gang Len Chang, HW Ho, and Yue Liu: Variation based online travel time prediction using clustered neural networks. In 2008 11th International IEEE Conference on Intelligent Transportation Systems, pages 85–90. IEEE, 2008.
7. P. Backus, M. Janakiram, S. Mowzoon, C. Runger and A. Bhargava, "Factory cycle-time prediction with a data-mining approach," in IEEE Transactions on Semiconductor Manufacturing, vol. 19, no. 2, pp. 252-258, May 2006, doi: 10.1109/TSM.2006.873400.
8. Yu, Tingting and Hongxia Cai: The prediction of the man-hour in aircraft as- sembly based on support vector machine particle swarm optimization. Journal of Aerospace Technology and Management, 7(1):19–30, 2015.
9. Fließ S, Dyck S, Schmelter M et al. (2015) Kundenaktivitäten in Dienstleistungsprozessen-die Sicht der Konsumenten. In: Kundenintegration und Leistungslehre. Springer, pp 181–204
10. Lantow, B. & Klaus, K. Analysis of Long-Term Personal Service Processes Using Dictionary-Based Text Classification. In *Human Centred Intelligent Systems* (pp. 77-87). Springer, Singapore.
11. Mirizio, G.G., Nunes, R.S.M., Vargas, D.A. et al. Time-of-Day Effects on Short-Duration Maximal Exercise Performance. Sci Rep 10, 9485 (2020).