# Predicting the Contractual Full-Time Equivalent Percentage Using XGBoost

Knut Håkon Grini, knut.grini@ssb.no

Stine Bakke, stine.bakke@ssb.no

**Abstract**

Statistics Norway receives monthly registry data concerning jobs, remunerations, and taxes (including payroll taxes) through a-ordningen, which was established in 2015. This is a mandatory electronic reporting system for anyone who has employees or who pays salary, pension, or other benefits.

A-ordningen also includes information on contractual full-time equivalent (FTE) percentage, i.e., the number of working hours stated in the employment contract relative to the number of working hours corresponding to a full-time job. However, the reporting of this variable is both questionable and incomplete, especially for hourly-paid jobs. To improve data quality in this respect, we have applied a method using XGBoost (eXtreme Gradient Boosting), which is a machine-learning algorithm based on gradient boosted decision trees.

All observations in a given month are initially run through various checks to detect outliers. Observations still considered valid after the outlier detections are used to fit and test an XGBoost model for the FTE of monthly salary, including socio-economic factors such as, e.g., sex, age, education, occupation, industry, and monthly salary as predictor variables. This model is subsequently used to predict the FTE percentage, for the remaining invalid observations.

In this paper, we give an overview of the XGBoost method and describe its application in predicting the contractual FTE percentage in cases where the reported values are either missing or considered erroneous.

**Keywords**: A-ordningen, contractual full-time equivalent percentage, prediction, registry data, machine-learning, XGBoost

## Introduction

Statistics Norway receives monthly registry data concerning jobs, remunerations, and taxes (including payroll taxes) through a-ordningen. This is a mandatory electronic reporting system for anyone who has employees or who pays salary, pension, or other benefits.

A-ordningen includes information on the contractual full-time equivalent (FTE) percentage, i.e., the number of working hours stated in the employment contract relative to the number of working hours corresponding to a full-time job. However, the reporting of this variable is both questionable and incomplete in many cases, especially for hourly-paid jobs. The FTE is essential for the register-based employment statistics, but also for publishing statistics on earnings, with major users such as the Technical Reporting Committee on the Income Settlements, research and official studies institutes, employee and employer organisations and Eurostat. To improve the quality of FTE, we have applied an imputation method using XGBoost (eXtreme Gradient Boosting), which is a machine-learning algorithm based on gradient boosted decision trees.

In this paper, we give an overview of the XGBoost method and describe its application in predicting the contractual FTE percentage in cases where the reported values are either missing or considered erroneous. Note that we are not describing the model itself but rather presenting its utility and establishing an understanding of why it is helpful.

## A-ordningen

### A-ordningen in general

A-ordningen was established in 2015 and is a collaborative system between Statistics Norway, the Norwegian Tax Authority and the Norwegian Labour and Welfare Administration. A-ordningen provides a digital service, where the information is submitted electronically every month via the employer’s payroll system to the Tax Administration Shared Services Agency, which administers a-ordningen.

Anyone who has employees or who pays salary, pension or other benefits, such as social security, must submit an electronic form called an a-melding. This is regulated by law (A-opplysnings act) with a lower threshold amount of NOK 1 000 per year[[1]](#footnote-1). The a-melding does not cover self-employed.

In summary, A-ordningen is a collaborative system, while the A-melding is an electronic form submitted to this system. Every year, 250 000 employers submit information concerning 4.8 million employees and pensioners via the a-ordning. E.g. type of employment, start and end data for the job, contractual full-time equivalent (FTE) percentage, salary and remuneration in cash and kind, and pension. Employers are also required to report information about some deductions, employer’s national insurance contribution and financial activity tax. More information about a-ordningen is available at the Norwegian Tax Authority’s website (Skatteetaten.no, 2019).

### Submission validation checks

The data submitted via the payroll systems are automatically checked for errors and omissions. Feedback is given to the employers almost instantly and includes information about discrepancies, if such are found, or if no discrepancies only a receipt is provided. The checks are classified according to the degree of severity which varies from rejection of the a-melding caused by a severe error, to a guideline feedback urging the reporter to check for potential errors in the delivered a-melding. Data are forwarded to Statistics Norway the 10th of every month which is 5 days after the deadline for reporting on the month expired. Upon arrival at Statistics Norway the data are processed, controlled and verified automatically. In addition to the controls integrated in the data delivery procedure, several automatic controls are implemented in our production system e.g. a routine for verifying jobs as active or not. Some employers may submit information about employees that are not active during the reference period, for instance seasonal workers or temporary staff where an end date of the job has not been reported. The verification is based on information on wages, duration of leave, benefits related to sickness, maternity leave etc.

Employers may submit replacements or corrections for earlier periods. It’s important to note that Statistics Norway bases most of its statistics from this source, on data received up to 2 months after the reference period. The main reasons being that the coverage is better (as we include delayed forms), and that corrections, for the most part, are submitted the month just following the reference month.

An important part of data preparation is the extensive matching with existing registers in Statistics Norway, most importantly population, business and education, but several others are also utilized.

### The statistical unit

The basic statistical unit in a-ordningen is jobs (or employment by local kind of activity), linked to each employee and remuneration. Usually the unit reported through the a-melding is a person working in a given establishment (a locally delimited functional unit). But the employer is free to report the job as two (or more) part-time jobs, which is a relatively common practice especially in the healthcare and social services. In our production however, part-time jobs in the same establishment are aggregated to one job, but not across establishments. If a person has more than one job, the main job is classified according to a set of criteria, including the type of job, working hours and so forth.

## Agreed working hours – and the problem

The quality of a-ordningen is generally considered very good and far better than previous sources for statistical purposes. A-ordningen provides a better coverage of part-time jobs as well as more accurate data at the individual level, than the main source up to and including 2014 (the State Register of Employers and Employees administered by the Norwegian Labour and Welfare Organization, NAV). There are several reasons why the quality has improved. With a-ordningen we have a coherent set of rules that force more accurate reporting. Information about working conditions and earnings now come from the same source, and not different sources as before. Coherence in the reporting between jobs and wages provides a better opportunity for follow-up and validity checks, and thus achieving better quality. In the case of the earnings statistics specifically, a-ordningen replaced a sample survey and therefore benefitted substantial improvements due to the same reasons listed above as well as elimination of the sampling error. However, the reporting of the contractual full-time equivalent (FTE) percentage is both questionable and incomplete in many cases, especially for hourly-paid jobs.

More specifically, working hours are not reported for many hourly-paid jobs, or it seems like a high proportion use default values such as 0 or 100. The latter is problematic as it complicates the process of identifying the incorrect values of the FTE percentages. After more than four years of collecting data, our experiences indicate that when the working hours and wages are inconsistent – the wages are usually reported correctly.

Since data about compensation are used for calculation of taxes for each individual, and also utilized in case management for calculating rights in connection with sick leave and other social benefits, the incentives to report correctly and timely are strong. The logic being that the more critical the use, the higher the barriers are for incorrect information. There is a possibility for each employer to access a report that sums up what has been reported either per month or for any other chosen period. This has been extensively utilized by many to check the consistency between reported data and data as noted in ledgers and accounting systems. This service gives two basic improvements as viewed by Statistics Norway, first and foremost quality, but also openness – we achieve a more explicit communication and understanding of how and what the information is used for.

As wages and working hours are correlated – we want to utilize the information we have on earnings and develop an imputation method for predicting the FTE percentage.

However, we want to emphasise that even though earnings are more often correct, we do find discrepancies: As mentioned previously, it is relatively easy to remove jobs that are not active in a given month, as we can identify who has not received a wage in the specified month. But, when employers submit information about employees that are not related to the current period we are looking at – a problem arises. This might be “right” according to the Norwegian Tax Authority, who are not concerned with which month the payment is linked to but when remuneration changes hands. For statistical purposes on the other hand, this is problematic as connecting the wage and FTE percentage becomes challenging.

## Preparing data – cleaning and identifying

When aiming to make use of a machine-learning (ML) algorithm to improve data it is important to secure a sufficient level of quality in the training dataset. If the model is given poor quality data, it will proceed and reproduce this substandard in the data you wish to improve, which we can agree is counterproductive.

In the following, we focus on some of the most significant quality challenges and how they are overcome. In this process we also identify observations in the data that need improvement.

There are three main stages for identifying faulty data. For the rest of the paper we will use the terms qualified and disqualified observations to establish a distinction between observations to learn from and observations that need to be discarded and predicted. Disqualification does not immediately imply that the observation is wrong, but indicates it is not representative and is excluded from the training dataset. Before the qualification routine can be applied however, a routine for tidying the reported data is necessary. Note that these procedures, both the qualification routine and the tidying, are important parts of preparing the data for the final statistical products, not just for the machine-learning.

### Cleaning/tidying the reported data

#### Smoothening the wage rate

In order to identify the fluctuations in the agreed FTE percentage we do not want the basic monthly earnings in the relevant period to contain payments for other periods. As mentioned in chapter 3, this might occur in the data. To correct for payments in the current month that are either connected to other periods or corrections for previous reporting, we smooth the wage/hourly rate for both fixed salary and pay by the hour, hereby called wage rate. Smoothing implies using the wage rate from the following month instead of the current one. The wage rate for the current month is compared to the wage rates of previous and next month in order to identify possible irregularities, in the form of either a peak or a drop in the wage rate for the current month. To prevent the use of extreme values from other periods, it should be noted that before we replace the wage rate with next month’s, we check that the number of working hours for employments is within a reasonable range. This range corresponds to an FTE percentage lower than 120.

#### Exclusion of extreme values

Inspired by an already existing method in the earnings statistics where outliers/extreme values are identified and new values for earnings are imputed – we decided to exclude some outliers in this stage of the process. Outliers that we believe should not be replicated/reproduced by XGBoost, due to high wage rates, are therefore excluded from the model but included in the final statistics, after a final outlier check. Creating indicator variables for having a high wage rate, a high salary, and either of those two, respectively.

* 1. An hourly wage rate of above 1,000 NOK (105 EURO) and/or
	2. Agreed wage rate (excluding fixed additional allowances) above 162,500 NOK (16,780 EURO) per month

#### Corrections of the FTE percentage

The last thing we do before the qualification routine is to revise some of the reporting of the FTE percentage:

1. If the FTE percentage is missing or equal to 0 the following occurs:
	* The FTE percentage is set equal to the number of paid hours for hourly paid employees
	* The FTE percentage is set equal to 100 for employees with fixed regular pay
2. If the FTE percentage surpass 120 it’s revised down to 120

### The qualification routine

The data is now ready to be processed in the qualification routine, which consists of the following checks:

Check 1:

* This check is conducted for hourly paid employees and aims at establishing if the employment should keep its reported agreed hours, or if paid hours should be used instead.
* A ratio model is utilized to identify extreme observations, the model is based on an iterative weighted linear regression with stratification. If the relationship between paid hours and the reported agreed hours is extreme/an outlier, the observation is marked as an extreme – but all observations continue to the next check before possibly being disqualified.

Check 2:

* A lower wage threshold is established through known lower boundaries. In Norway there is no minimum wage or all-encompassing rules that regulate such a boundary. However, our experience indicates that the following limits are reasonable for the FTE wage:
	+ Less than 6,000 NOK (620 EURO) for apprentices
	+ Less than 12,000 NOK (1,240 EURO) for workers younger than 18 years
	+ Less than 18,000 NOK (1,860 EURO) for all other
* Lower and upper limits for hourly paid employees are set
	+ Lower limits are currently set for hourly paid employees at 100 NOK (10 EURO)
	+ Upper limits are set at 1000 NOK (105 EURO)

The disqualified observations are set aside for imputation, while the qualified observations continue to the next checks.

Check 3:

* Is an iterative linear regression model that checks for outliers in the relation between the dependent variable, the logarithm of the FTE agreed/basic earnings and the logarithm of agreed/basic earnings, and other explanatory variables such as occupation and sector. Observations that are judged to have too great a discrepancy are disqualified and imputed using prediction from the XGBoost model.

The results from the qualification routine are summarised in Table 1, showing that the proportion which will be used for learning, the ML-method is about 93 percent. While the proportion that are imputated using prediction by XGBoost is close to 7 percent.

Table 1. Qualification results from September 2018 (N=2,762,030 observations) per cent of total

|  |  |  |
| --- | --- | --- |
|  | Qualified | Disqualified[[2]](#footnote-2) |
| Employees with fixed regular pay | 65,0 | 3,7 |
| Employees with hourly wages | 27,5 | 3,8 |
| Totals | 92,5 | 7,5 |

Source: Statistics Norway

## Use of XGBoost

The basic needs or reasons for employing imputation using machine learning is to improve data quality. Either you wish to improve inaccurate or false data or predict partial non-existent data. For this reason, the training has to be based on the best data you can lay your hands on. Identifying or cornering in this crucial element can prove just as daunting a task as compiling and producing the actual statistics.

Why then bother, you could ask. It is the answer to this question that also presents itself as the strength of machine learning. As long as you nurture and update your model appropriately it will improve the production process and reduce the future workload and hopefully provide consistent predictions from one period to another. Investing in the future in other words.

There are several machine learning tools to choose from for imputation, that have their strengths and weaknesses. At any rate, the aim is always to choose a model that is best suited for solving the particular problem at hand. A small number of related models were initially viewed as relevant, amongst these were Random Forest, Gradient Boosting Decision Trees and Extreme Gradient Boosted Decision Trees (XGBoost), which was our final choice. The model itself is based on the fundamentals from so called decision trees.

As earlier stated this paper will not aim to describe this model, technically or mathematically, but rather present its utility. We therefore more or less treat the model itself as a black box and rather emphasize how to identify and establish the relevant training data, and thereby also test data, and presenting some consequences of making use of training data that are inadequate or substandard.

### Practical use

Of a total of 2,555,968 qualified observations 20 per cent are drawn from this total randomly and are to be used as dedicated test data. The 80 per cent left provide the training data.

When XGBoost is applied, several parameters are set and specifications of selected variables are constructed (Xgboost.readthedocs.io, 2019). Most importantly, variables that are going to be used for “growing trees”/establish patterns for predicting values are chosen:

Continual variables:

* Age
* Age squared
* Number of employees in local unit (categorized in ten groups)
* LOG(fixed earnings) (monthly earnings excluding variable allowances and bonuses)
* Actual reported FTE in per cent or calculated FTE (based on calculation from paid hours)

Categorical variables

* Gender
* Highest registered education (first digit ISCED)
* Industry according to NACE2007
* Occupation (first two digits ISCO2008)
* Apprenticeship
* Earning category (fixed monthly, hourly paid, other)

Even though one of the specifications for utilizing XGBoost is assigning a model, in our case a linear model, you do not specify the model itself only the variables that are to be used. Therefore, describing a singular model would be strictly wrong since many models are constructed, for increasingly improved fit, but describing it would also be outside the realm of this paper. Categorical variables are transformed into dummy variables. Since the chosen ML and others make use of several “trees” it will also employ covariance between the variables included.

When it comes to choosing variables for the ML some choices are obvious, while others not so much. The relations between earnings and agreed working hours are among the obvious, the same goes for industry, age, gender and occupation. Not so clear is the recalculation of basic earnings to LOG. XGBoost does not require this but seemed to produce better results compared to a non-transformed variable. The exclusion of bonuses and irregular payments is foremost apparent because these vary between months and do often represent remunerations for work done in earlier months, and do not improve quality of prediction. The size of the local unit does matter for certain industries and within these across some occupations, but not others. Amongst other points we find that larger local units in some industries employ a large number of jobs with small FTEs. This might be because these larger units are better equipped to utilize many small jobs with uncertain availability, such as students or people seeking shorter terms of employment.

Several other parameters have to be set, even though XGBoost comes with pre-set values that regulate how aggressive or conservative the ML will behave. Changing some settings may also adversely affect memory and/or speed of execution (Xgboost.readthedocs.io, 2019). Even though this is not within the frame of this paper, it is important to make it clear that adjustments on pre-sets should be done with at least some understanding of both data and the final aim for performance both actual and in prediction.

### Actual results

Let us assume that we serve the model with “bad” data, data that are faulty/unclean. A model that behaves appropriately would reproduce the same patterns in the target data that we want to correct, the ML assumes that bad data are correct, the assumption would be that a good model would reproduce a predictable amount of incorrect results just as it was given to learn from.

To put this into perspective we can compare the consequences of the aforementioned test with doing the same operation with clean data as the base for learning.

What we want to achieve with this is to highlight that the ML actually is able to reproduce the specific characteristics we want to improve, and of course that the results can be used to describe and understand the reality the statistics are aimed at addressing. In other words, we must not lose ourselves in a technical wizardry or inherent technical and mathematical functions, as if that was a goal in itself. Our main goal is to efficiently produce the best statistics possible that serves user needs. Thereby, the criteria for judging performance are the results: do the final results more efficiently describe reality than the existing methods applied in production of statistics.

In the following we wish to present some illustrations to summarize some important results from the application of either inadequate data compared with adequate data. Our purpose was initially to improve faulty or missing data for agreed hours by making use of existing data, at the same time, this could come in conflict with existing statistics on earnings from the same source. Remember, information about earnings are utilized in this work, both for qualification and disqualification of observations or the lack thereof.

In diagram 1 we have presented a random subset of data where the results are based on data that are not vetted as good as they should be. The x-axis is FTE while the y-axis is earnings per full time equivalent, recalculated after prediction of new values for the FTE. Diagram 2 on the other hand presents a result that is much better. Both diagrams are based on the same sub-sample and have the same scale (please note the pattern of the reported values). It is also important to note that the comparison is done with actual data used in the publication of earnings statistics and has therefore passed through our regular production routine.

Diagram 1 Results after applying inaccurate data for learning

Source: Statistics Norway

In diagram 1, at least one observation has received input that puts it very high up on the earnings ladder. But most important is the fact that most of the predicted values on the x-axis, FTE, are bunched up below an FTE of 70 per cent. In diagram 2, this is changed, and the observations are more spread across the range of FTE.

Diagram 2 Results from a more accurate learning data set

Source: Statistics Norway

In general, the predicted values in diagram two are now more neatly ordered in the distribution of earnings and FTE, as a point of simple eyeball economics.

## Final output and the results

In the following we wish to present some of the results and try to explain and understand if and how the model helps us to improve the statistics. We also have to address the question: does the application of ML introduce new problems for the production or presentation of statistics?

One systematic trend in the training data is that paid hours used to estimate agreed hours tends to underestimate agreed hours. This can be the case because regular leave or lunch breaks are not included in the paid hours, or the employee was not present due to other reasons. This has to be refined in the training data and a final judgment is yet to be passed.

How the distribution of FTE has changed is the most important feature of study in this work, because it is the objective of utilizing ML in the first place. In table 2, we have presented the distribution of FTE by deciles and presented mean values in total and for each decile and the percentage of the sum of FTE.

The main summary is that the FTE after prediction has slightly more data at the lower end of the scale, the mean FTE is at 80 per cent after prediction compared to 82 per cent in the original data. While the total number of full-time jobs is approximately the same in both (the difference in distribution is significant, both in amount and statistically).

Table 2 Decile distribution of FTE in the earnings statistics before and after imputation[[3]](#footnote-3)

|  |  |  |
| --- | --- | --- |
|  | Average FTE | Percent of total FTE |
| Decile | Before prediction  | After prediction | Before prediction  | After prediction |
| **Total** |  **0,82**  |  **0,80**  | **100,0** | **100,0** |
| 1 |  0,16  |  0,12  | 2,0 | 1,5 |
| 2 |  0,43  |  0,35  | 5,3 | 4,4 |
| 3 |  0,68  |  0,60  | 8,3 | 7,5 |
| 4 |  0,92  |  0,90  | 11,2 | 11,3 |
| 5 |  1,00  |  1,00  | 12,2 | 12,5 |
| 6 |  1,00  |  1,00  | 12,2 | 12,5 |
| 7 |  1,00  |  1,00  | 12,2 | 12,5 |
| 8 |  1,00  |  1,00  | 12,2 | 12,5 |
| 9 |  1,00  |  1,00  | 12,2 | 12,5 |
| 10 |  1,02  |  1,02  | 12,4 | 12,7 |

Source: Statistics Norway

Furthermore, the distribution is significantly more unequal in the new predicted data due to more frequent low values, measured with a Gini-coefficient[[4]](#footnote-4), and the same is apparent if measured as skewedness or variance.

This would in most cases have to be viewed as problematic, and rightfully is, as for earnings statistics we suddenly have more money to distribute on less working hours and for employment statistics we might be underestimating agreed working hours.

We have still not been able to asses whether we have further possibilities in the prequalification or in improving the model specifications, or if these predicted values actually reveal problems in the existing statistics and thereby the data. A combination of the two mentioned hypotheticals might be just as correct.

### Passing judgment

Experience so far has highlighted several interesting experiences that we would like to share and not the least disclose to further discussion and notice for applying ML. The ML model we have chosen is fast and seems to be fairly easy to implement. ML consistently produces a variation in data that is more consistent with what we are able to produce with other methods we have applied.

XGBoost is sensitive to the quality of the learning data and especially to the existence of extreme values/outliers. One does not have to feed the mentioned outliers into the model but can choose to hold the most extreme observations outside of the model, which we have chosen to do. The mentioned extremes will afterwards be added to the data just not utilized in the learning data. The reason for this being that for most of these extremes have been found in industries and occupations that are not represented in the data that need prediction through ML.

When first set up our choice of ML does need continual supervision, in other words we have to check both parameters and preparation of the training data. If that is not done the result would be an increasing substandard quality of both results from the ML and in our case inefficient data-preparation. Our long-term goal is that this preparation and imputation process should be an integrated part of our production of annual and quarterly labour market and earnings statistics.

## Summary

The current application of ML seems, at the moment, to be slightly biased toward smaller percentages of FTE. However, this tendency of bias is not across the board, and we are quite optimistic. We have tested several methods for selecting training data and several sets of variables and transformations.

Even though our challenge with improving data on FTE seems small in the context of having such a great data source as the a-melding at our disposal, it is important to try to meet ever increasing demands from our users. While working on using ML, we have identified possible improvements to our current production system. In perspective we are only adjusting data that we otherwise would impute through more crude methods or even worse, discard. An improvement to FTE, through the use of ML, has already made it possible to produce more statistics on the distribution of agreed working hours and FTE.

Last but not least, the use of the current data source has been a short story so far, only four years, with a steep learning curve. Through this work we have identified several possibilities for improvements but also for distributing even more detailed and accurate statistics about the labour market and earnings in the future. Hopefully, through this first test of ML, we might gather the courage to utilize ML in other instances.

## References

Brownlee, J. (2019). *A Gentle Introduction to XGBoost for Applied Machine Learning.* [online] MachineLearning Mastery. Available at: <https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning>

GitHub. (2019). *dmlc/xgboost*. [online] Available at: <https://github.com/dmlc/xgboost>

Pafka, S. (2019). *Benchmarking Random Forest Implementations | Data Science Los Angeles*. [online] DataScience.LA. Available at: <http://datascience.la/benchmarking-random-forest-implementations>

Skatteetaten.no. (2019). *About the a-ordning*. [online] Available at: https://www.skatteetaten.no/en/business-and-organisation/employer/the-a-melding/about-the-a-ordning/ [Accessed 26 Jun. 2019].

Xgboost.readthedocs.io. (2019). *XGBoost Documentation*. [online] Available at: <https://xgboost.readthedocs.io/en/latest/>

1. In some cases, for small enterprises that do not have a payroll system, the a-melding can be submitted online electronically via the governmental dialogue system *Altinn* [↑](#footnote-ref-1)
2. Reasons for disqualification differentiate as described above [↑](#footnote-ref-2)
3. Includes all observations, not only those imputed [↑](#footnote-ref-3)
4. The Gini-coefficient is a statistical measure of dispersion. A Gini coefficient of zero expresses perfect equality, where all values are the same. A Gini coefficient of 1 expresses maximal inequality among values. [↑](#footnote-ref-4)