Research Paper Business Analytics

The gender pay gap in the technology sector

Writer: Anne Jonker, master Student Business Analytics

Student number: 2528666

Start date: 01-06-2018

Finishing date: 01-11-2018

Supervisor: Mr. Bernard Zweers

Grade: 8.5 / 10

Research paper on gender pay gap in the technology industry

Anne Jonker - Master student Business Analytics, Vrije Universiteit Amsterdam - 2528666

***Abstract*— This paper investigated the features that could explain the salary and employee grades in the technology sector. The dataset contained 12.259 employees who operated globally and was collected from 2016-04 till 2018-05. The dataset only allowed to compare the different regions: EMEA, APAC and Americas due to insufficient data.**

The methods used to determine these factors were: Multiple Linear Regression, Random Forest and Extreme Gradient Boosting. Each of the methods had their parameters optimised trough grid search and were programmed in Python 3.6.

When only the averages are taken of the employees big differences in salary and employee grades were observed, similar to the literature. However when these features were corrected by other factors the difference diminished. The difference in salary is largely due to the skewed distribution of gender in the employee grades. There are more men in higher positions and therefore obtain a higher salary and employee grade. The overall best method in determining the salary and employee grades was making use of the Extreme Gradient Boosting method.

No hard conclusions can be drawn based on this dataset except for the fact that it is crucial to correct for age, functional level and other features in order to determine the gender gap.

I. INTRODUCTION

Reducing the gender pay gap is a high priority within developed countries. It is clear that equal work should result in equal pay, however it is believed that there is a so-called ”glass ceiling” or ”sticky floor” effect. Both of these terms refer to the fact that female employees are faced with an invisible and almost insurmountable barrier when it comes to increases in salary and promotion opportunities beyond a certain level, while their male co-workers are not hindered by such barriers. These effects and other literature will be further described in section II.

As such, for example Iceland has decided to enforce strong equal pay laws partially as a response to large-scale protests in 2016. The country already had equal pay laws in place but believed that if these laws were not sufficiently enforced the gap would remain, and as such the country vowed close the gender pay gap by 2022.

Furthermore, the UK recently published a list of companies where there was a gender pay gap present, in order to publicly shame and steer them towards offering equal pay. Many other countries are still struggling to determine how exactly to tackle this issue, but overall the consensus is that such a gap exists and that it is imperative to reduce it.

However, despite this wide-spread belief, academic literature does not offer a definitive insight into this subject due to different approaches to calculating such a gap. For example, in the UK, current practice dictates that the salaries of all male and female employees should be averaged to determine whether there is a difference at all. This neglects to take into account many other factors that could and perhaps should influence one’s salary. Therefore, their calculations result in figures up to 57%, indicating that for every euro a male employee earns, a female employee would only earn 57 cents.

Given that many of the methods used do neglect to take many possibly relevant factors into account, this paper aims to investigate what actually determines the salary of an employee and specifically whether gender is a significant contributor therein. Even if that turns out not to be the case, gender could underlie another variable; and as such those situations and practices should still be considered at least indirectly discriminatory. Furthermore, the different models (described in section IV) and results (described in section V) derived herein will then be compared and contrasted to conclude whether differences can be observed between the EMEA, the Americas and the APAC regions.

The data that is used in this paper comes from a single technology company which operates globally. The contents and transformations in this dataset is thoroughly explained in section III.

II. LITERATURE

There have already been many publications devoted to the subject of gender gaps. However, the factors that possibly impact to the salary of an employee and are studied in said publications depend heavily on the data available to the respective researchers.

A large-scale study conducted by [Plantenga et al., 2006] shows that the variation in determining the actual figure of one’s salary can lead to conflicting conclusions. There are, however, a few parallels that can be drawn that represent consistencies between these studies.

First of all, when a random sample of the population is used, there is a tendency for the resulting gender pay gap to be higher. However, when the sample is taken among employees that are just starting their careers, the gap narrows. In other words, the gap widens as employees work for their companies for a longer duration. Implicitly, this also means that the gender gap varies with age.

Secondly, the gender pay gap in the public sector tends to be smaller than the gender pay gap within the private sector. One issue with this observation is that the public sector is to a large degree dominated by women, which could contribute to a distorted view of reality if these results are generalised. [Arulampalam et al., 2007]

Lastly, it appears that marital status is of some influence. The gender pay gap is smaller for single employees and

wider for married couples. This is another contentious subject due to the fact that the likelihood of a married couple having children is larger then that of a single parent. As a result, women often take maternity leave and are overall more likely to take on the responsibility of taking care of their children, at least in traditional households, which continue to be prevalent throughout western societies. In turn, this can result in lower working hours available to female employees that have children, and thus results in decreased chances of salary increases and promotions. [Waldfogel, 1998b] This situation could be remedied by applying state policies such as paid maternity leave. Countries who have these policies tend to have smaller cumulative family wage gaps than those who do not.

According to [Hedija et al., 2015] one contributor to the gender pay gap is the gender of one's manager. It concluded that *"women in middle management in comparison to their male counterparts have a lower tendency to apply wage discrimination against women. The presence of a female head of department led to a decrease in the gender pay gap by almost 7 percentage points."*

Education and the educational level of the parents are other factors that could significantly influence the salary [Davis-Kean, 2005]. In the past, less women would participate and complete forms of tertiary education than men would. This resulted in an increase in the gender pay gap. However, this situation is trending towards women participating more and more in every form of education [OECD, 2015]. However, the fields that male and female students graduate in are not equally distributed. According to [Ayalon, 2003] participation by female students in the mathematical and science-related fields is far lower than in humanistic fields, comparatively. [Ashenfelter and Mooney, 1968] shows that expected salaries are significantly higher for those in mathematical and science-related than in the humanistic fields.

The gender gap is decreasing over time in most countries [Oostendorp, 2004], but vary in their rate of change. An international meta study conducted by [Weichselbaumer and Winter-Ebmer, 2005] states that this is due to better labour market endowments for women, due to changes in policy.

The article of [Petit, 2007] showed a significant hiring discrimination in France towards young women (aged 25 and under), specifically within high-skilled jobs.

Another study conducted by [Stuhlmacher and Walters, 1999] investigated the differences in outcomes of negotiations with gender as the dependent variable. It concluded that "*although the overall difference in outcomes between men and women was small, none of these hypothesized moderators or several exploratory moderators reversed or eliminated this effect.*"

The article of [Small et al., 2007] investigated this subject more in-depth and concluded that the circumstances surrounding and setting in which a negation takes place are also very important. Women tend to find it more intimidating to start the negotiation if an opportunity to do so arises. This could result in a lower-placed employee position or lower overall salary due to the fact that women would concede sooner during these negotiations.

The article of [Krings and Olivares, 2007] also showed that ethnicity is a factor in the decision whether or not to hire an employee. Especially groups generally disliked by the population of the specific countries were much less likely to be hired.

III. DATA

The data used in this paper contains information on 12,259 employees who work for one large international company operating in the technology sector and were followed for three years (from 2016-04 to 2018-05). Due to privacy reasons most of the data was hashed but remains usable by the various methods employed. The salary of the employees was converted to euros, using the exchange rates of the same month the data was recorded, to compare the different salaries among the different countries.

The remaining features were one-hot coded [Sutter et al., 2002]. One-hot coding converts the different categorical features to booleans with either a 1 as a value if it is true, or a 0 if it is false. An example of this is that gender was split up into two new features with "gender = M" and "gender = F". The original gender feature was then removed, to avoid collinearity (features that can almost be fully explained by another feature [Belsley et al., 2005]). The reason for this is that many methods cannot handle categorical data, but all can handle one-hot coded datasets. Unfortunately the data on martial status, ethnicity and education levels were not provided and therefore could not be tested on, even tough the literature suggests that this could influence salary, as previously mentioned in section II.

An important feature is the employee grade. This grade is used in many large firms to compare different positions within the company. The investigated company uses the following employee grades, which range from 1 to 22: 1 is the lowest grade and 22 the highest. To determine the employee grade of an employee many factors are involved: the number of responsibilities, education level, the number of other employees supervised and additional schooling.

Another feature is the Business unit, which refers to the field in which an employee is working. This can be for example: Human Resources, Legal, Research and Development, and so on.

The Functional level is the feature referring to which role an employee has. This can be for example: Manager, Director, or Senior Consultant.

The Business unit and Functional level features were all hashed and only limited to the first level, for reasons similar to the region limitation as will be discussed hereafter: there would be insufficient data to draw any meaningful conclusions if there are divided further. The full list of included features are shown in Appendix II.

The first step in preparing the data was to clean it. Rows with many empty fields were removed together with rows

were the gender or salary was not provided. Other employee data was removed if the data could not possibly be correct. An example was a yearly salary of €6,000,000,000.-, or birth dates that have not occurred yet. This resulted in a final data set for this study that covered 10.874 employees.

The data is split up in four different datasets. The first dataset contains all observations and will be referred to as Global. The other three datasets are based on three different regions: Americas, Europe Middle East and Africa (EMEA), and Asia Pacific (APAC). Unfortunately it was not possible to delve any deeper into these regions due to insufficient data, as this would lead to statistical problems. The distribution of the number of employees in the different regions is shown in Table I. The table shows that the Global dataset is dominated by the EMEA region, which should be taken into account when drawing conclusions.

Region	% of Global
APAC	25.7%
EMEA	58.3%
Americas	15.9%

TABLE I

DISTRIBUTION OF REGIONS

	Gender			Average age		
	F	M	% Diff.	F	M	% Diff.
Global	38%	62%	23%	30.59	30.84	0.8%
APAC	28%	72%	43%	28.96	29.55	2.0%
EMEA	41%	59%	17%	31.22	31.48	0.8%
Americas	44%	56%	11%	30.14	31.05	2.9%

TABLE II

GENDER AND AVERAGE AGE PER REGION

Table II, obtained after cleaning the datasets, shows that more male employees are employed in every region. The biggest difference is observed in the APAC region. Contrastingly, the average age difference is overall relatively small. This would suggest that the age distribution in the company is distributed relatively equally.

	Average employee grade			Average yearly salary in €		
	F	M	% Diff.	F	M	% Diff.
Global	9.46	10.43	9.3%	23.340	35.017	33.4%
APAC	9.65	10.14	4.9%	14.245	19.800	28.1%
EMEA	9.46	10.59	10.7%	26.733	43.169	38.1%
Americas	9.28	10.44	11.1%	21.223	35.064	39.5%

TABLE III

EMPLOYEE GRADE AND AVERAGE YEARLY SALARY IN € PER REGION

The differences in Table III appear to be substantial. The employee grade, wherein the biggest difference is shown the Americas region (11.1%), would mean that a female employee would be around 10% lower on the employee grade scale. The average yearly salaries in euros is especially interesting. Looking at these numbers alone one could observe that female workers make far less than their male co-workers. At a global level, the difference is 33.4% and it gets as high as 39.5% in the Americas.

	Corrected time in position		
	F	M	% Diff.
Global	0.74	0.55	33.4%
APAC	0.51	0.43	14.4%
EMEA	0.83	0.61	26.5%
Americas	0.67	0.58	12.7%

TABLE IV

CORRECTED TIME IN POSITION AND PROMOTION PER REGION

Table IV shows the corrected time in one's position. The corrected values are based on number of promotions and corrected for employees who left the company early. It shows that a male employee on average would be promoted earlier than a female employee.

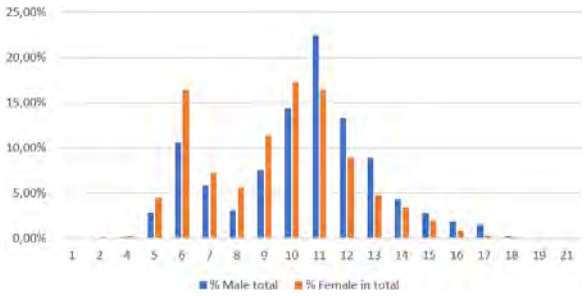

Fig. 1. Gender distribution employee grade Global

Figure 1 shows the distribution of gender in employee grades. It shows that the graph is skewed in the sense that up until grade 11, women are over-represented in the lower employee grades and men in the higher employee grades.

The numbers will be processed in the next section where three distinct methods will be used to see if these numbers are indeed correct and if the pay gap is as large if the previous numbers suggested.

IV. METHODS

To determine the various factors of the height of the salary three different methods are used. The programming has been done in Python 3.6, using the packages of SKlearn. Parameter optimising is conducted on each of the different methods. In the subsections the parameters will be explained together with their optimal settings.

A. RMSE

To compare the performance of the different regression methods used, the Root Mean Squared Error (RMSE) is taken of each of the models. The RMSE is shown in equation 1, where $\hat{\theta}$ is the predicted value of an observation and θ the actual value from the test set. The difference between the observation and actual value is then squared. Afterwards the expectation of the squared difference is calculated and finally has its root taken. The lower the RMSE, the better the model was able to predict the salary or employee grade.

$$RMSE = \sqrt{MSE(\hat{\theta})} = \sqrt{\mathbb{E}((\hat{\theta} - \theta)^2)} \qquad (1)$$

B. Training and testing

To train and test the different methods, the datasets were randomly split into 80% train data and 20% test data. To ensure that the models did not train on different sets, the same sets were used in the three models. The 80%/20% split is considered common practice in the data science field. To enhance the quality of the models, a seven-fold cross-validation is applied on each of the models [Pedregosa et al., 2011].

C. Multiple Linear Regression (MLR)

The Multiple Linear Regression (MLR) is a method that tries to build a formula using p predictors (x_{pi}) to predict a (dependent) target variable Y_i, while minimising the error (ε_i). Each predictor has a coefficient (β_i), which can be positive or negative, to correct the predictors feature importance. The β_0 is the intercept term, which is a constant. ε_i is the error term which is the difference between the predicted and the actual value. This results in equation 2.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \varepsilon_i \tag{2}$$

If (β_i) is large, then the influence of the predictor (x_{pi}) is also large in determining the target variable [Aiken et al., 2003]. This method is one of the most common ones in the data science field and therefore used as the first method to create a baseline.

An example of a possible model build by the MLR on the dataset is shown in equation 3. This equation suggests that to calculate the salary only three features are significant (above the $\alpha = 0.05$ threshold). The FTE, employee grade of the employee and Gender of employee, which is in this case only relevant if the employee is a woman.

If we want to calculate the predicted salary of two example employees, described in V, we obtain the equations 4, 5 for employees 1 and 2. What would be remarkable to see is that even tough employee 2 works more hours (higher FTE) and is in a higher employee grade than employee 1, it still earns less salary. The difference lies in the fact that employee 2 has a true value for Gender = F which means that employee 2 is female, while employee 1 is male.

$$\begin{aligned}\text{Salary} = 2548.02 + 0.9(\text{FTE}) - 150(\text{Gender = F}) \\ +0.2(\text{employee grade})\end{aligned} \tag{3}$$

$$\text{Salary}_1 = 2548.02 + 0.9(0.9) - 150(0) + 0.2(10) = 2550.83 \tag{4}$$

$$\text{Salary}_2 = 2548.02 + 0.9(1) - 150(1) + 0.2(11) = 2401.12 \tag{5}$$

Employee	FTE	Gender = F	Employee grade
1	0.9	0	10
2	1	1	11

TABLE V
EXAMPLE OF TWO EMPLOYEE SALARY CALCULATION

D. Ensemble learning methods

The Random Forest and XGB are both ensemble learning methods. Before the Random Forest and XGB are discussed in more detail, the ensemble learning method is explained first.

An ensemble learning method is one where many classifiers are generated and finally have their results aggregated [Liaw et al., 2002]. There are two methods to accomplish this, either tough bagging [Breiman, 1996] or boosting [Schapire et al., 1998] of classification trees.

In the bagging method, each new tree is independent of previously constructed trees and is built trough a bootstrap sample of the given dataset. The prediction is determined by a majority vote.

In the boosting method, each new tree allocates additional weight to points that were not predicted correctly by previous trees. The prediction is unlike the bagging method determined by a weighted vote.

Bagging and boosting both provide higher stability for the model by reducing variance. The differences are that boosting tries to reduce the bias, while bagging does not. Bagging might solve the over-fitting problem, while boosting could only increase this. Therefore it is important to compare tree models that either perform boosting or bagging [Bauer and Kohavi, 1999].

E. Random Forest

The Random Forest algorithm is an ensemble learning method, with one additional important feature. The Random Forest changes the bagging method in two ways. The first manner is that instead of using the same bootstrap sample of the data for every tree, a new bootstrap is taken every time a new tree is built. Secondly it changes how a tree is built entirely. In the traditional way, nodes are split using the best split with all features taken into account. The Random Forest makes a subset out of the features at the node that is going to be split. The predictions that are used in the voting process are determined in the following manner.

At the construction of a new tree (z) a set of features, $f(z)$, is randomly sampled. Then a regressor, $d = R(f(z))$, is trained to predict the most likely position of the target variable relative to z [Cootes et al., 2012]. These predictions are then used in a vote to determine the best position in an accumulator array V. The variable v represents the degree of confidence in the prediction. This results in equation 6.

$$V(z + d) \rightarrow V(z + d) + v \tag{6}$$

An example of this on the dataset can be found in section IV-G. The article of [Breiman, 2001] shows that it outperforms most other classifying algorithms and is more robust in combating noise and overfitting.

The parameters that were optimised were the number of estimators (the trees in the forest) and the maximum depth of the tree. The optimal parameters were obtained through a grid search. The Random Forest Regression was used in this paper to determine the salary and employee grade.

F. Extreme Gradient Boosting Machine

The Extreme Gradient Boosting Machine (XGB) is a fairly new method which can be used (like the Random Forest algorithm) for classification and regression purposes and uses tree structures [Chen and Guestrin, 2016]. The article explains the mathematical approach and scoring functions in depth.

Like the Random Forest, the XGB is also an ensemble method. It adds predictors and improves upon the previous models in various ways. The difference with the Random Forest is that it fits the model to new residuals of the previous prediction after each iteration followed by minimising the loss function after the latest prediction has been added. This is in contrast with the Random Forest, which changes the weights given to the classifiers. The name of this method is clarified by the fact that the loss function updates the model using a gradient descent. The difference between normal Gradient Boosting and XGB is that XGB has an additional custom regularisation term in the objective function. An example on how this method works is described in section IV-G.

To measure the impact of a feature on the model the Shapley Additive Explanation (SHAP) is used. SHAP represents how much the model was affected by adding the feature to the constructed trees. The paper of [Lundberg and Lee, 2017] explains the mathematical workings and the axioms it satisfies.

In this paper the XGB regression is used to determine the salary and employee grade, since both of these variables are continuous.

The three different methods will work on four datasets. First the entire dataset, Global, is analysed, followed by each region: APAC, EMEA, and Americas.

G. Example of Random Forest and XGB

This section is used to clarify the two methods using an example dataset shown in Table VI.

Employee	FTE	Gender = F	Age	Employee grade
1	0.9	0	33	10
2	1	1	28	11
3	1	1	51	15
4	0.5	0	32	8
5	1	1	30	??

TABLE VI
EXAMPLE DATASET

The Random Forest follows these steps:

1) The first tree has to be constructed.
2) It chooses a random number of employees from the dataset to base predictions on (this can also be the same employee multiple times).
3) This is the dataset on which only this tree will be built.
4) From all possible features (FTE, Gender = M, age, and so on) it chooses a random number of features, and decides, based on these features, how the tree is split into new nodes, using the best split possible. The next nodes are split using the same feature sample and best split method. The best split is determined by the information gain in the ID3 matrix. The paper of [Ferri et al., 2002] discusses this more in detail.
5) This process is repeated until the number of specified trees is built and a forest appears.
6) When the forest is constructed the trees that predicted the employee grades correctly are taken and used as the final model. This is done by counting the number of trees that predicted correctly.

Lets assume for the first tree the employees 1, 2, 3, 4 are randomly selected together with all the features. This resulted in Figure 2.

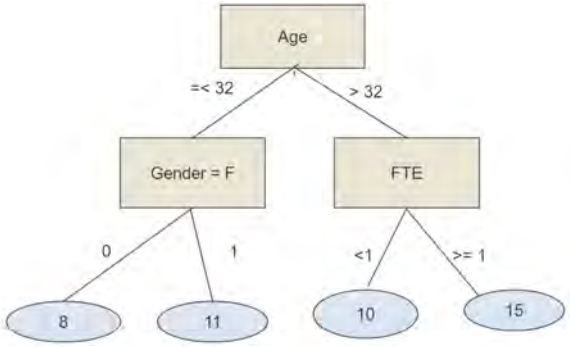

Fig. 2. Random Forest: Tree 1, predicting employee grades

The next tree only employees 1, 3, 4 are randomly selected and only the features age and FTE. This resulted in Figure 3.

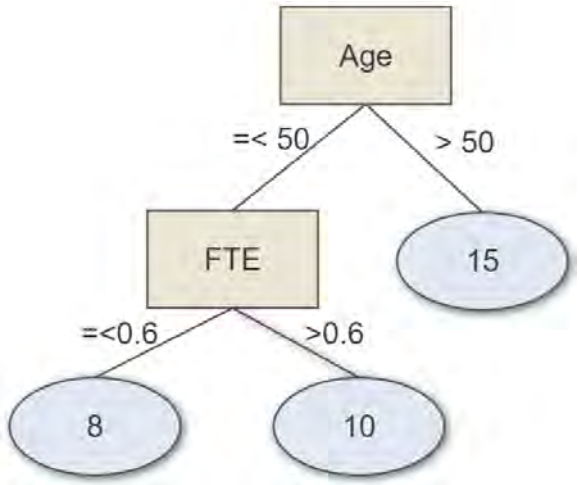

Fig. 3. Random Forest: Tree 2, predicting employee grades

This process repeats for 750 more times (parameter setting of number of estimators). Now lets assume that the second tree is the most accurate based on the other trees in the Random Forest. Then the model would predict that the employee grade of employee 5, is 10.

The same trees can be constructed trough the XGB method. The difference however is in the sampling of the

employees. If we assume that Tree 1 did not predict the employee grade for employees 6, 7 and 8 (not shown but random examples), these employees will have a higher probability to be selected in the random sample in constructing a new tree. Therefore it learns from its "mistakes" in the past and adapts the model. The problem with this is the sensitivity for strong outliers, since it will always try to correct for this value.

V. RESULTS

The results start with an overview of each model with their optimal parameter optimisation. This is followed by the quality of each model and ends with the feature importance outcomes.

A. Parameter optimisation

The multiple linear regression has a limited amount of parameters that could be tuned. The first parameter is the tolerance based on which a feature is added or removed from the model, of which the optimum was found to be 0.05.

The second parameter represents whether the MLR should work with the so called step-up or step-down method. The step-up method refers to starting with an empty model and adding the feature with the highest R^2-value. The model is executed and in each iterative step the next highest R^2-value is added until the subsequent feature is no longer significant. A feature can not be deleted from the model once it has been added.

The step-down method is similar, but instead of starting with an empty model, it starts with all features added. The features with the lowest significance gets removed and the model is evaluated again. This step is repeated until all features in the model are significant. The problem that could occur with this method is that it is more likely that features remain in the model that are not necessarily needed in the model [Bendel and Afifi, 1977]. After testing both of the methods the step-up method was ultimately used.

For the Random Forest method the only parameter that was optimised was the number of trees. Due to computation time and memory usage this was set to 750 trees. The other parameter is the maximum depth of each tree, which was not optimised in this paper due to computation limitations.

Table VII shows the optimal parameter settings for the XGB method with regards to salary, and Table VIII with regards to employee grade.

	Sub sample ratio	Max. depth	Number of estimators
Global	0.8	18	204
APAC	0.8	17	201
EMEA	0.8	17	202
Americas	0.8	16	198

TABLE VII

OPTIMAL PARAMETERS FOUND XGB REGARDING SALARY

	Sub sample ratio	Max. depth	Number of estimators
Global	0.8	18	205
APAC	0.8	17	203
EMEA	0.8	17	201
Americas	0.8	17	199

TABLE VIII

OPTIMAL PARAMETERS FOUND XGB REGARDING EMPLOYEE GRADE

B. Quality of models

To compare the performance of the three models the RMSE is used. Table IX shows the error margins and the R^2 of the MLR. These models were used to determine the factors that influence salary.

	MLR	MLR(R^2)	Random Forest	XGB
Global	25547.84	0.66	3900.74	3801.80
APAC	18001.56	0.65	2090.40	2216.34
EMEA	29922.99	0.69	4403.87	4288.52
Americas	19677.26	0.78	4531.53	4441.77

TABLE IX

RMSE COMPARISONS ON THE THREE DIFFERENT MODELS ON SALARY

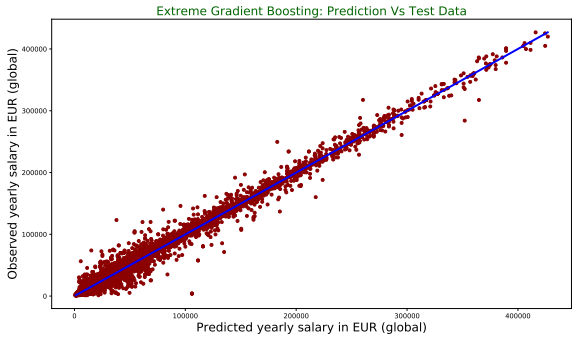

Fig. 4. XGB: fit on Global salary prediction

Figure 4 shows the fit of the of the XGB model to the Global test set.

Table X shows the quality of the models on determining employee grades. Figure 5 shows the fit of the XGB to the global data set.

	MLR	MLR(R^2)	Random Forest	XGB
Global	1.68	0.63	0.73	0.73
APAC	1.52	0.56	0.88	0.88
EMEA	1.68	0.68	0.75	0.72
Americas	1.46	0.71	0.57	0.56

TABLE X

RMSE COMPARISONS ON THE THREE DIFFERENT MODELS ON EMPLOYEE GRADE

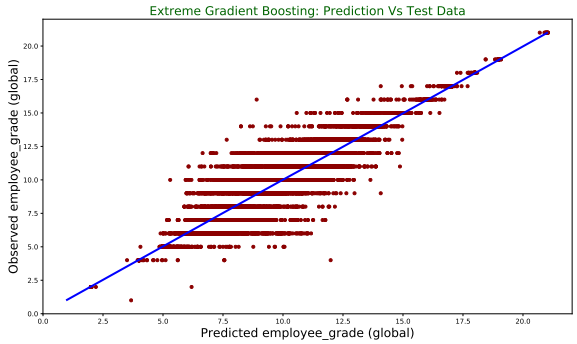

Fig. 5. XGB: fit on Global employee grade prediction

The MLR method appears to be very unsuccessful in determining salary and employee grades. The RMSE is much higher than those of the Random Forest and XGB. The R^2 values, which represents the percentage of the response variable variation that is explained by a linear model, is however relatively high [Baarda et al., 2011]. But due to the much larger RMSE the MLR is no longer considered to be a good method for these datasets and will not be discussed in the remaining parts of this paper.

The results of the Random Forest and XGB are similar with respect to the RMSE. The XGB appears to outperform the Random Forest when it comes to salary, except when only the APAC region is considered.

C. Results of the models

The results of the Random Forest can be found in the Appendix, since only the XGB is discussed due to the lowest RMSE.

1) Salary prediction: The Random Forest and XGB models were able to predict the salary quite well, as shown in Table IX. The features that were important in determining salary were consistent between the Random Forest and XGB methods. It is noteworthy that in both models the employee grade was the most significant factor across all datasets. Figure 6 shows the impact of this feature on the model for the Global dataset. This outcome was very similar to the models when applied to the other datasets, which can be observed in Appendix section I.

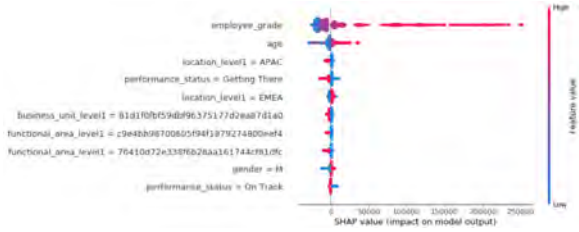

Fig. 6. XGB: Feature importance salary in Global dataset

2) Employee grade prediction: Since salary, target bonus and position grade all presented a large collinearity with the Employee grade variable, these were removed from the dataset. The same models (with different parameter settings due to optimisation) were executed using the same train and test sets as discussed in the previous section. Figure 7 shows the top ten features that impacted the model for the global dataset. The fact that the functional levels and age were still more important than gender is interesting, however a slight impact is measured between the gender of the manager and gender of the employee. Figures 8, 9, 10 show the different impacts of the features in the models. However, the features that impact the models are roughly the same, but differ in the significance thereof. The APAC region, however, showed no impact on the model at all.

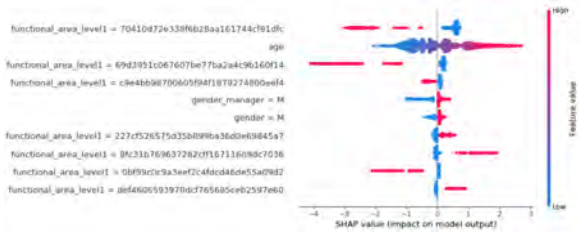

Fig. 7. XGB: Feature importance employee grade in Global dataset

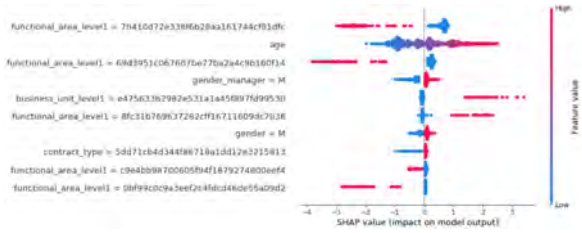

Fig. 8. XGB: Feature importance employee grade in EMEA dataset

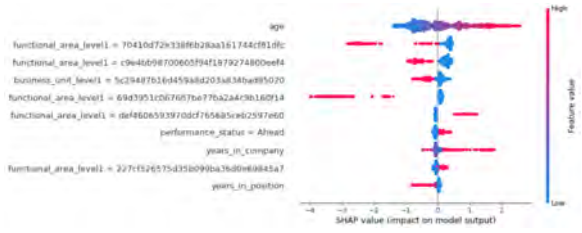

Fig. 9. XGB: Feature importance employee grade in APAC dataset

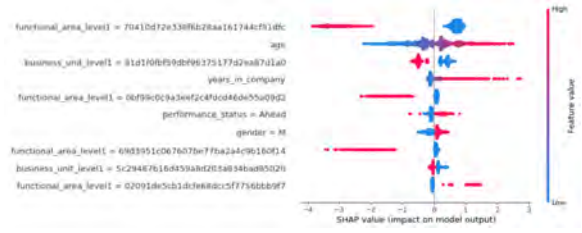

Fig. 10. XGB: Feature importance employee grade in Americas dataset

VI. DISCUSSION

Similar to the papers as previously discussed in the literature section II a large difference can be observed between the employee grades and yearly salaries depending on one's

gender. As shown in Table III, the differences range from 28.1% up to 39.5%. This is consistent with what the current literature stated and reported [OECD, 2015].

The key problem lies in determining the manner in which the gap is calculated. If factors such as age, function levels, employee grades, and many others are not taken into account, the true reality is obscured. Therefore it is crucial to carefully establish a calculation for the gender pay gap. Many conclusions have been drawn and presented as facts in political debates, which in turn were based on very biased calculations.

The dataset used in this paper is very far from perfect. Some features were missing like the education level, ethnicity and marital status. Furthermore it was only specified for relatively large regions. Cultural differences could influence the outcome of any research when it comes to the impact that gender has, and could be clarified if country-specific data were to be made available. Furthermore, only one company in the technology sector was used for this research, which could also have lead to biased conclusions. It could feasibly be possible that this company employs additional policies that aim to reduce pay gaps that other companies may not, or vice versa.

The three methods used showed that there are much better ways to determine the features that influence the salary and employee grades than those typically used. MLR, in the end, could not predict accurately enough to be considered, however Random Forest and XGB could. In hindsight the models would perform even better if the salaries would have been classified in different bins [Leow and Li, 2001]. The concept of binning is to change the exact salary values into different bins. An example of this would be to aggregate all salaries between 0 and 10.000 euros, 10.001 until 20.000, and so on. This way the models could operate as classifying methods instead of regression methods. The models tried to predict values that could never be predicted precisely. Therefore the RMSE was still present and was sometimes even relatively high. However, in that case the question of how the number and size of these bins should be determined should first be addressed. However, such binning probably would have reduced the over-fitting and enhanced the performance of the XGB even more.

There appears to be some influence of the gender of the manager and the gender of the employee, however not as much as one's age and functional level. The distribution of gender within the employee grades showed that men occupy, on average, higher positions than women. This automatically results in higher salaries and employee grades, and should not be disregarded when trying to draw conclusions from this paper, as this affects the models and therefore their outcomes. The same observation can be made when examining the feature importance graphics derived from the XGB. The model was only affected if either the employee was male or the manager was male.

It was possible to remove the features of age and functional level to see if the gender feature becomes a more important feature in the impact of the model regarding employee grade determination. The problem with this is that due to the lack of features gender might appear to be far more important than it really is. The methods include gender more often than not, not necessarily because it is a genuine explanatory variable, but rather because it is used in place of missing variables.

When examining only the raw statistics that resulted from the Random Forest and XGB, gender does not seem to be of great importance. However the problem is that only looking at numbers could lead to wrong conclusions. An unbiased and critical human interpretation is always necessary. As previously described in the literature section it was shown that there is a difference between genders when it comes to negotiating. Men tend to negotiate more and better than woman. This could lead to a difference in salary and employee grade in general. Men and woman are simply not the same, however by stating that men should therefore earn more is inherently wrong. If it turns out that there exists actual and systemic discrimination based on one's gender, this should be eradicated altogether.

However, in my personal opinion, I don't believe that the gender gap will ever be fully closed. Not due to discrimination but due to biology. The vast majority of women have the desire to have children and this will continue to influence their career path. Maternity leave is crucial in the reintegration process of women in the workforce and countries with similar policies have a higher working participation of women [Waldfogel, 1998a]. However, If policies would change to where women would earn more based on the fact that they are women, this is still discrimination. The goal in decreasing the gender pay gap is not that women earn the same amount as men, but to level the playing field in the work environment. Equal work should result in equal pay.

The final problem in drawing conclusions from this paper has been that its results cannot be extrapolated to represent the gender gap in general. Only one company in the private technology sector has been investigated, where from furthermore data was missing on features that may prove to be key indicators.

Therefore it will always be very hard to calculate the gender pay gap to any degree of precision. However, just determining plain averages within companies is a very bad idea. Comparing the different countries is also harder than it appeared at first. Countries have different policies in place considering maternity leave, unemployment and their education systems. Therefore, data has to be corrected for all of this in order to provide a fair comparison between countries, and to do so would warrant an entirely new study altogether. Statistics and complex models are great resources, however a clear, critical, and unbiased human mind also needs to weigh their results carefully.

VII. CONCLUSION

Determining the gender pay gap and employee grades depends on the manner in which it is calculated and the datasets that are used. Large differences can be observed when only the average salaries of men and women are

examined. However, when other features are taken into account to correct for any differences, the gender gap does not appear as large. The main feature that influences the height of one's salary turns out to be their employee grade within a company.

Therefore, by removing features that have a high collinearity with employee grade, the following features were observed to be significant in determining the employee grade. The most important features are the functional levels and age. This is followed by the gender of the manager and gender of the employee. Still, this does not necessarily mean that there is active discrimination going on, based on gender. It should however be investigated how it is possible that men typically occupy higher positions than women. The answer to this would be the most probable explanation for the gender pay gap, based on the dataset that was used.

The methods can still be improved by feeding the methods with more data. Additionally, finding a way to correctly bin the different salaries would further improve the accuracy of the models.

In the end, no definite conclusions can be made on the height and existence of the gender pay gap.

VIII. ACKNOWLEDGEMENT

This study could not have been conducted without the help of a number of people. Firstly, I would like to thank my supervisor drs. Bernand Zweers. It was a tremendous pleasure working together, especially in determining the required methods, and I am thankful for all of the provided feedback. Furthermore I would like to thank Mr. Dirk Jonker (Crunchr) for supplying the dataset and Mrs. Rianne Kaptein (Crunchr) for the counselling on the dataset. Mr. Tjalling Otter was there to correct the many grammar flaws and overall structuring of the English sentences. Finally Mr. Bram Jonker in helping to clarify my findings and final presentation.

REFERENCES

[Aiken et al., 2003] Aiken, L. S., West, S. G., and Pitts, S. C. (2003). Multiple linear regression. *Handbook of psychology*, pages 481–507.

[Arulampalam et al., 2007] Arulampalam, W., Booth, A. L., and Bryan, M. L. (2007). Is there a glass ceiling over europe? exploring the gender pay gap across the wage distribution. *ILR Review*, 60(2):163–186.

[Ashenfelter and Mooney, 1968] Ashenfelter, O. and Mooney, J. D. (1968). Graduate education, ability, and earnings. *The Review of Economics and Statistics*, pages 78–86.

[Ayalon, 2003] Ayalon, H. (2003). Women and men go to university: Mathematical background and gender differences in choice of field in higher education. *Sex Roles*, 48(5-6):277–290.

[Baarda et al., 2011] Baarda, B., de Goede, M. P., and van Dijkum, C. (2011). *Basisboek statistiek met SPSS*. Noordhoff.

[Bauer and Kohavi, 1999] Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, 36(1-2):105–139.

[Belsley et al., 2005] Belsley, D. A., Kuh, E., and Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons.

[Bendel and Afifi, 1977] Bendel, R. B. and Afifi, A. A. (1977). Comparison of stopping rules in forward stepwise regression. *Journal of the American Statistical association*, 72(357):46–53.

[Breiman, 1996] Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

[Cootes et al., 2012] Cootes, T. F., Ionita, M. C., Lindner, C., and Sauer, P. (2012). Robust and accurate shape model fitting using random forest regression voting. In *European Conference on Computer Vision*, pages 278–291. Springer.

[Davis-Kean, 2005] Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: the indirect role of parental expectations and the home environment. *Journal of family psychology*, 19(2):294.

[Ferri et al., 2002] Ferri, C., Flach, P., and Hernández-Orallo, J. (2002). Learning decision trees using the area under the roc curve. In *ICML*, volume 2, pages 139–146.

[Hedija et al., 2015] Hedija, V. et al. (2015). The effect of female managers on gender wage differences. *Prague Economic Papers*, 24(1):38–59.

[Krings and Olivares, 2007] Krings, F. and Olivares, J. (2007). At the doorstep to employment: Discrimination against immigrants as a function of applicant ethnicity, job type, and raters' prejudice. *International Journal of Psychology*, 42(6):406–417.

[Leow and Li, 2001] Leow, W. K. and Li, R. (2001). Adaptive binning and dissimilarity measure for image retrieval and classification. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE.

[Liaw et al., 2002] Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.

[OECD, 2015] OECD (2015). *The ABC of gender equality in education: Aptitude, behaviour, confidence*. OECD Publishing.

[Oostendorp, 2004] Oostendorp, R. (2004). *Globalization and the gender wage gap*. The World Bank.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

[Petit, 2007] Petit, P. (2007). The effects of age and family constraints on gender hiring discrimination: A field experiment in the french financial sector. *Labour Economics*, 14(3):371–391.

[Plantenga et al., 2006] Plantenga, J., Remery, C., et al. (2006). The gender pay gap. origins and policy responses. a comparative review of thirty european countries. *Synthèse du rapport pour la Commission Européenne, Equality Unit. http://www. reteparioppportunita. it/Rete_Pari_ Opportunita/UserFiles/news/report_ pay_gap_economic_experts. pdf ENFANTS, INTERRUPTIONS DACTIVITÉ DES FEMMES ET ÉCART DE SALAIRE ENTRE LES SEXES*.

[Schapire et al., 1998] Schapire, R. E., Freund, Y., Bartlett, P., Lee, W. S., et al. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686.

[Small et al., 2007] Small, D. A., Gelfand, M., Babcock, L., and Gettman, H. (2007). Who goes to the bargaining table? the influence of gender and framing on the initiation of negotiation. *Journal of personality and social psychology*, 93(4):600.

[Stuhlmacher and Walters, 1999] Stuhlmacher, A. F. and Walters, A. E. (1999). Gender differences in negotiation outcome: A meta-analysis. *Personnel Psychology*, 52(3):653–677.

[Sutter et al., 2002] Sutter, G., Todorovich, E., López-Buedo, S., and Boemo, E. (2002). Low-power fsms in fpga: Encoding alternatives. In *International Workshop on Power and Timing Modeling, Optimization and Simulation*, pages 363–370. Springer.

[Waldfogel, 1998a] Waldfogel, J. (1998a). The family gap for young women in the united states and britain: Can maternity leave make a difference? *Journal of labor economics*, 16(3):505–545.

[Waldfogel, 1998b] Waldfogel, J. (1998b). Understanding the" family gap" in pay for women with children. *Journal of economic Perspectives*, 12(1):137–156.

[Weichselbaumer and Winter-Ebmer, 2005] Weichselbaumer, D. and Winter-Ebmer, R. (2005). A meta-analysis of the international gender wage gap. *Journal of Economic Surveys*, 19(3):479–511.

APPENDIX I
ADDITIONAL CHARTS AND TABLES

A. *Global*

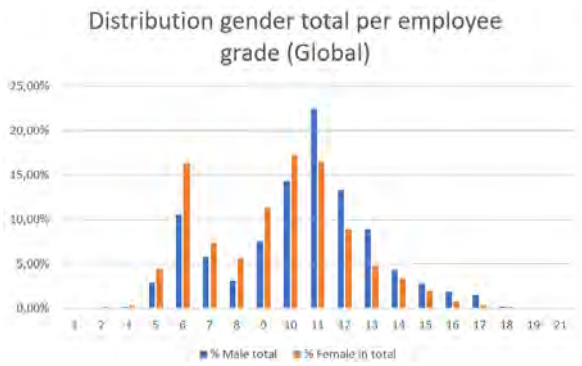

Fig. 11. Gender distribution employee grade in the Global dataset

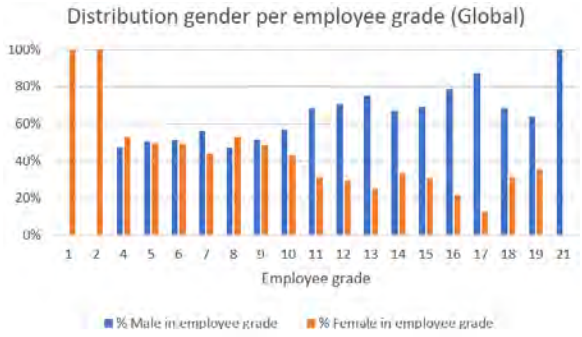

Fig. 12. Gender distribution per employee grade in the Global dataset

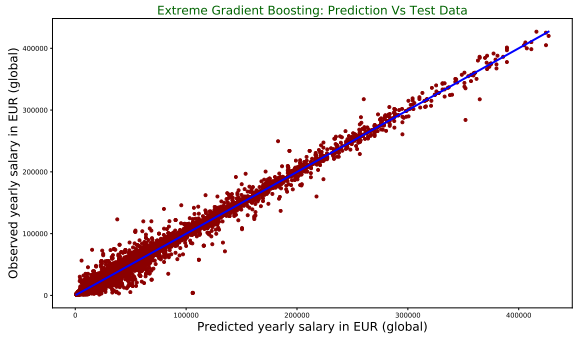

Fig. 13. XGB: fit on salary in the Global dataset

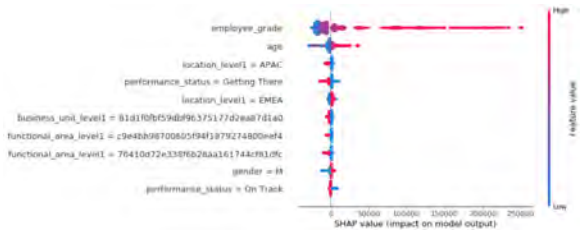

Fig. 14. XGB: feature importance on salary in the Global dataset

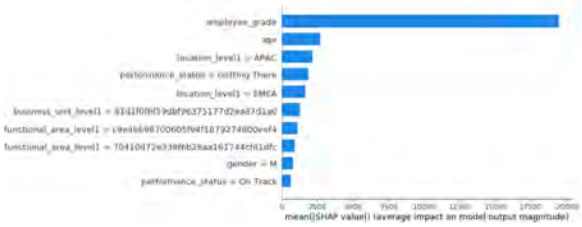

Fig. 15. XGB: Mean feature importance on salary in the Global dataset

Feature	Importance
Employee grade	0.78
Age	0.04
Years in company	0.02
Years in position	0.01
Performance status = Getting There	0.01

TABLE XI
RANDOM FOREST: IMPORTANCE ON SALARY IN THE GLOBAL DATASET

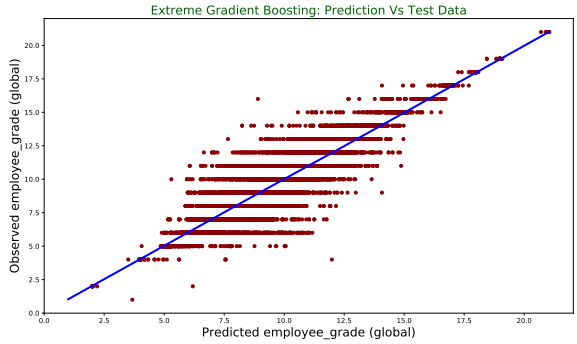

Fig. 16. XGB: fit on employee grade in the Global dataset

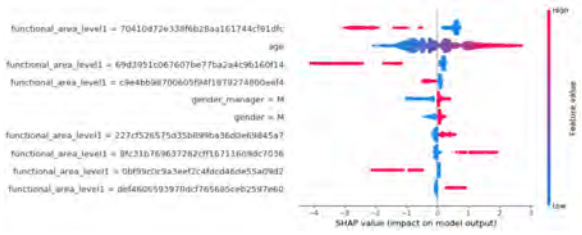

Fig. 17. XGB: feature importance on employee grade in the Global dataset

Fig. 18. XGB: Mean feature importance on employee grade in the Global dataset

Feature	Importance
Functional area level1 = 70410d72e338f6b28aa161744cf81dfc	0.26
Age	0.19
Functional area level1 = 69d3951c067607be77ba2a4c9b160f14	0.12
Years in company	0.05
Functional area level1 = 8fc31b769637282cff16711609dc7036	0.03

TABLE XII

RANDOM FOREST: IMPORTANCE ON EMPLOYEE GRADE IN THE GLOBAL DATASET

B. EMEA

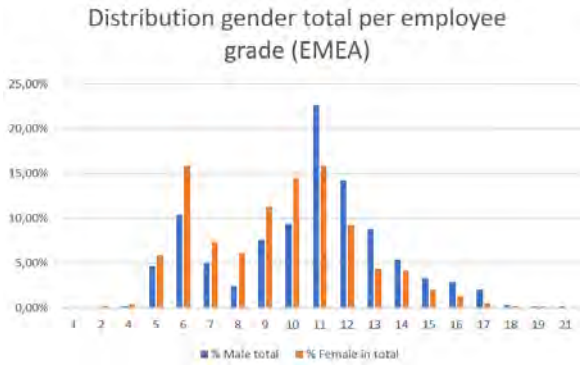

Fig. 19. Gender distribution employee grade in the EMEA dataset

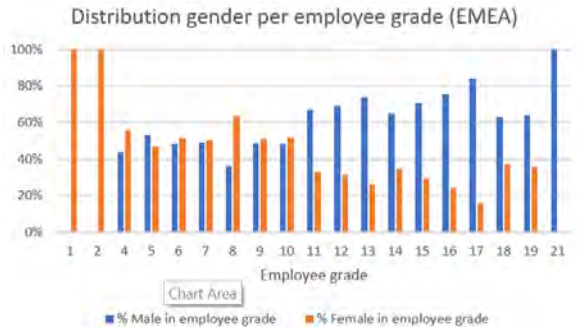

Fig. 20. Gender distribution per employee grade in the EMEA dataset

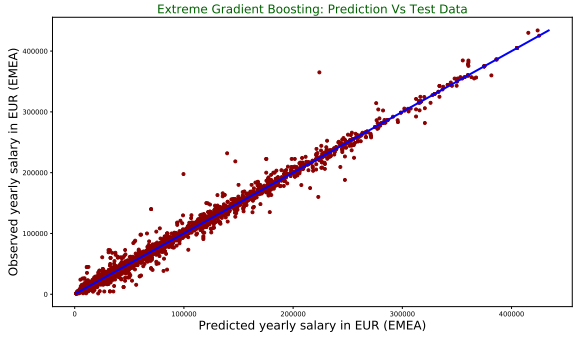

Fig. 21. XGB: fit on salary in the EMEA dataset

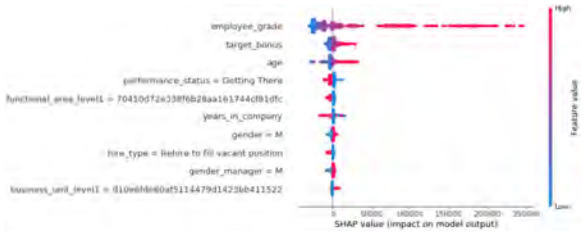

Fig. 22. XGB: feature importance on salary in the EMEA dataset

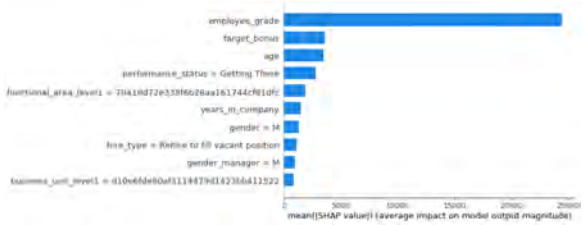

Fig. 23. XGB: Mean feature importance on salary in the EMEA dataset

Feature	Importance
Employee grade	0.79
Age	0.04
Target bonus	0.03
Years in company	0.02
performance status = Getting There	0.01

TABLE XIII

RANDOM FOREST: IMPORTANCE ON SALARY IN THE EMEA DATASET

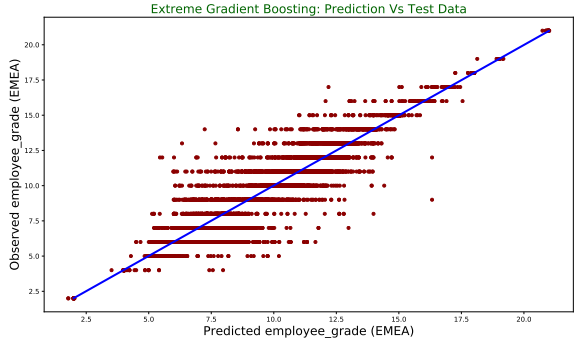

Fig. 24. XGB: fit on employee grade in the EMEA dataset

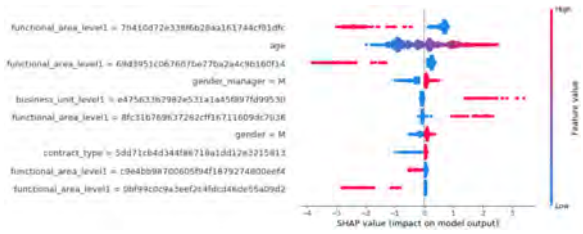

Fig. 25. XGB: feature importance on employee grade in the EMEA dataset

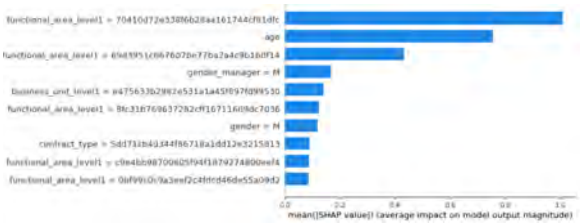

Fig. 26. XGB: Mean feature importance on employee grade in the EMEA dataset

Feature	Importance
Functional area level1 = 70410d72e338f6b28aa161744cf81dfc	0.26
Functional area level1 = 69d3951c067607be77ba2a4c9b160f14	0.18
Age	0.17
Years in company	0.05
Functional area level1 = 8fc31b769637282cff16711609dc7036	0.04

TABLE XIV

RANDOM FOREST: IMPORTANCE ON EMPLOYEE GRADE IN THE EMEA DATASET

C. APAC

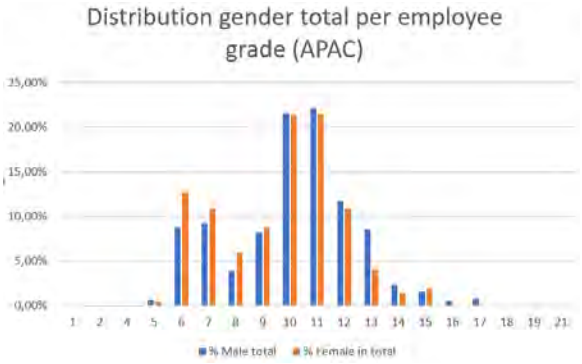

Fig. 27. Gender distribution employee grade in the APAC dataset

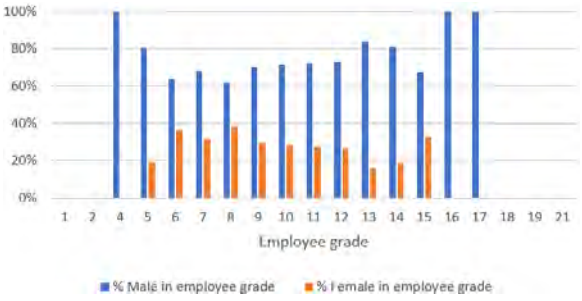

Fig. 28. Gender distribution per employee grade in the APAC dataset

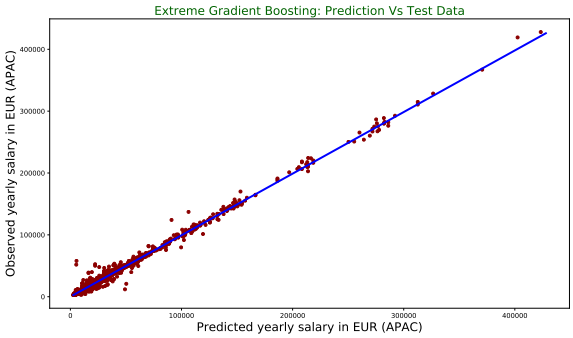

Fig. 29. XGB: fit on salary in the APAC dataset

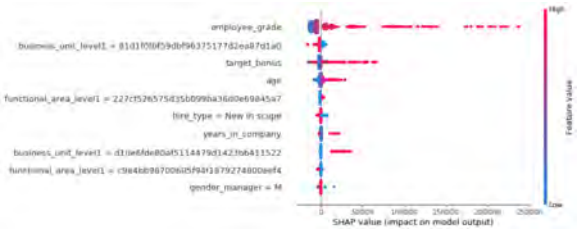

Fig. 30. XGB: feature importance on salary in the APAC dataset

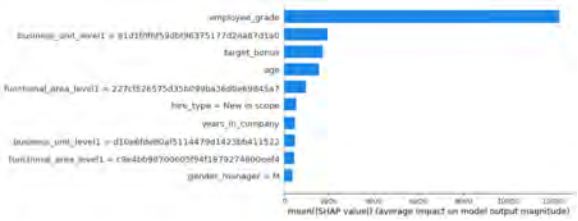

Fig. 31. XGB: Mean feature importance on salary in the APAC dataset

Feature	Importance
Employee grade	0.75
Target bonus	0.11
Age	0.02
Business unit level1 = 81d1f0fbf59dbf96375177d2ea87d1a0	0.02
Functional area level1 = f8285afedf7b10d55ca7656420723ab9	0.02

TABLE XV

RANDOM FOREST: IMPORTANCE ON SALARY IN THE APAC DATASET

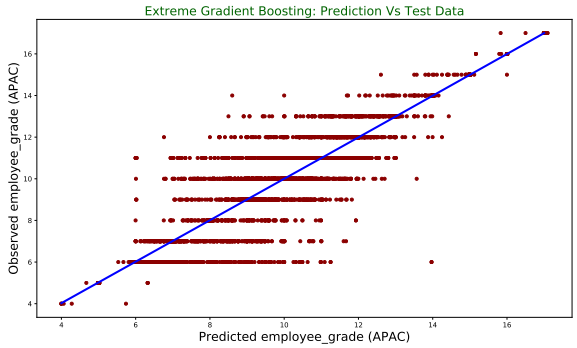

Fig. 32. XGB: fit on employee grade in the APAC dataset

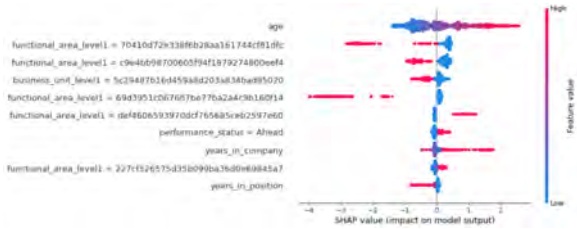

Fig. 33. XGB: feature importance on employee grade in the APAC dataset

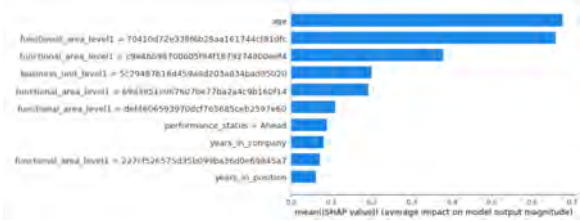

Fig. 34. XGB: Mean feature importance on employee grade in the APAC dataset

Feature	Importance
Age	0.22
Functional area level1 = 70410d72e338f6b28aa161744cf81dfc	0.21
Functional area level1 = 69d3951c067607be77ba2a4c9b160f14	0.09
Functional area level1 = c9e4bb98700605f94f1879274800eef4	0.08
Years in company	0.06

TABLE XVI

RANDOM FOREST: IMPORTANCE ON EMPLOYEE GRADE IN THE APAC DATASET

D. Americas

Fig. 35. Gender distribution employee grade in the Americas dataset

Fig. 36. Gender distribution per employee grade in the Americas dataset

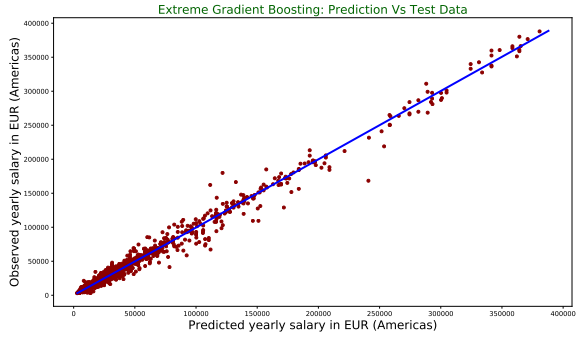

Fig. 37. XGB: fit on salary in the Americas dataset

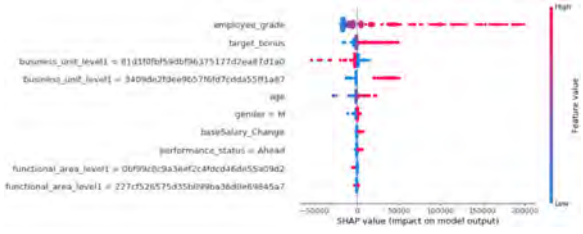

Fig. 38. XGB: feature importance on salary in the Americas dataset

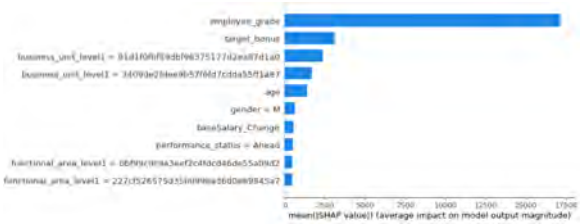

Fig. 39. XGB: Mean feature importance on salary in the Americas dataset

Feature	Importance
Employee grade	0.77
Business unit level1 = 3409de2fdee9b57f6fd7cdda55ff1a87	0.11
Age	0.03
Business unit level1 = 81d1f0fbf59dbf96375177d2ea87d1a0	0.03
Years in company	0.01

TABLE XVII

RANDOM FOREST: IMPORTANCE ON SALARY IN THE AMERICAS DATASET

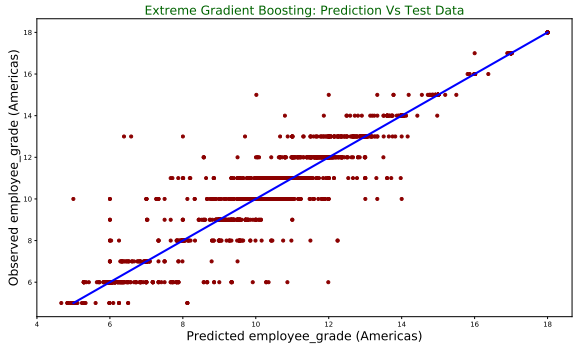

Fig. 40. XGB: fit on employee grade in the Americas dataset

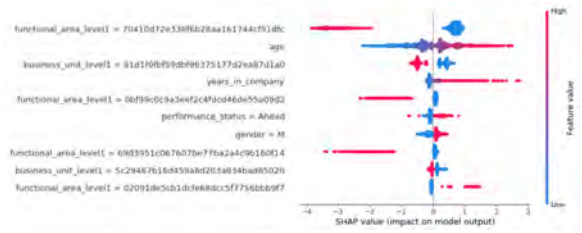

Fig. 41. XGB: feature importance on employee grade in the Americas dataset

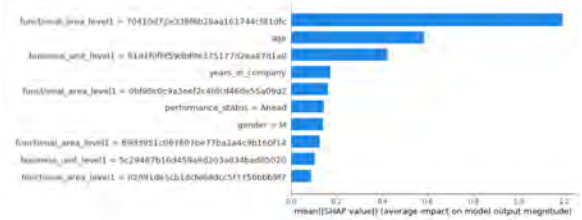

Fig. 42. XGB: Mean feature importance on employee grade in the Americas dataset

Feature	Importance
Functional area level1 = 70410d72e338f6b28aa161744cf81dfc	0.37
Age	0.19
Business unit level1 = 81d1f0fbf59dbf96375177d2ea87d1a0	0.07
Years in company	0.06
Functional area level1 = 69d3951c067607be77ba2a4c9b160f14	0.04

TABLE XVIII

RANDOM FOREST: IMPORTANCE ON EMPLOYEE GRADE IN THE AMERICAS DATASET

APPENDIX II
FEATURE NAMES AND MEANING

Feature	Explanation
Age	The age of the employee
Business unit level	The different divisions in the company and has three levels: It start with a general department (HR, Legal). Level two is a more specific part in this department and finally specific job description
Contract type	The type of contract an employee has. This can be for example a temporary, fixed or other types of contracts.
Employee grade	The employee grade determines the importance of an employee. The grade is determined by the number of responsibilities, education level, previous job experience and so on.
FTE	The amount of hours worked by one employee in a year on a full-time basis. 1 FTE = 2.080 hours in a year
Functional level	The function position of an employee. This can be head of a department, managing a sub division and so on.
Gender	Gender of the employee
Gender manager	Gender of the manager supervising the employee
Hiring status	The reason why an employee was hired
Location level	Location in which an employee is working has three levels. Region, country, city
Performance status	How the employee is performing: Getting there, on schedule, Ahead.
Position grade	The position grade is the level in the company an employee is in. It ranges from 1 (highest (CEO)) to 8 (new intern). This grade is highly correlated with the employee grade.
Solid line	The solid line variable is the employee_id of the manager of the manager of this employee. Trough iterative steps it was possible to determine the entire structure of the company.
Target bonus	The bonus an employee receives at the end of the year when certain targets are met. This is a multiplier of the yearly salary
Years in company	Number of years the employee has been working in the company
Years in position	Number of years the employee has been working in the current position

TABLE XIX
FEATURE NAMES AND MEANING

APPENDIX III
EMPLOYEE GRADE DISTRIBUTIONS

	Global		EMEA		Americas		APAC	
Employee grade	% Male in total	% Female in total	% Male in total	% Female in total	% Male in total	% Female in total	% Male in total	% Female in total
1	0.00%	0.02%	0.00%	0.01%	0.00%	0.06%	0.00%	0.00%
2	0.00%	0.14%	0.00%	0.22%	0.00%	0.05%	0.00%	0.00%
4	0.18%	0.31%	0.27%	0.48%	0.00%	0.07%	0.09%	0.00%
5	2.87%	4.46%	4.67%	5.90%	0.59%	3.83%	0.61%	0.37%
6	10.57%	16.39%	10.45%	15.96%	14.66%	21.72%	8.81%	12.68%
7	5.83%	7.28%	5.06%	7.41%	1.70%	3.05%	9.25%	10.93%
8	3.13%	5.65%	2.46%	6.13%	4.28%	3.71%	3.81%	5.97%
9	7.54%	11.42%	7.59%	11.34%	5.92%	14.43%	8.22%	8.77%
10	14.34%	17.31%	9.44%	14.50%	18.27%	22.60%	21.59%	21.45%
11	22.45%	16.49%	22.64%	15.93%	22.33%	13.22%	22.16%	21.49%
12	13.31%	8.94%	14.26%	9.24%	12.85%	5.87%	11.76%	10.92%
13	8.94%	4.78%	8.86%	4.45%	10.13%	6.55%	8.52%	4.13%
14	4.31%	3.41%	5.44%	4.17%	3.90%	2.89%	2.39%	1.41%
15	2.77%	1.98%	3.39%	2.02%	2.92%	1.95%	1.54%	1.88%
16	1.87%	0.82%	2.83%	1.30%	0.98%	0.00%	0.50%	0.00%
17	1.52%	0.35%	2.04%	0.55%	1.14%	0.01%	0.74%	0.00%
18	0.21%	0.15%	0.29%	0.25%	0.32%	0.00%	0.00%	0.00%
19	0.10%	0.09%	0.18%	0.14%	0.00%	0.00%	0.00%	0.00%
21	0.07%	0.00%	0.12%	0.00%	0.00%	0.00%	0.00%	0.00%

TABLE XX
EMPLOYEE GRADES IN THE VARIOUS REGION DATASETS