How much is a Data Science job candidate worth?

By Laura Gascue

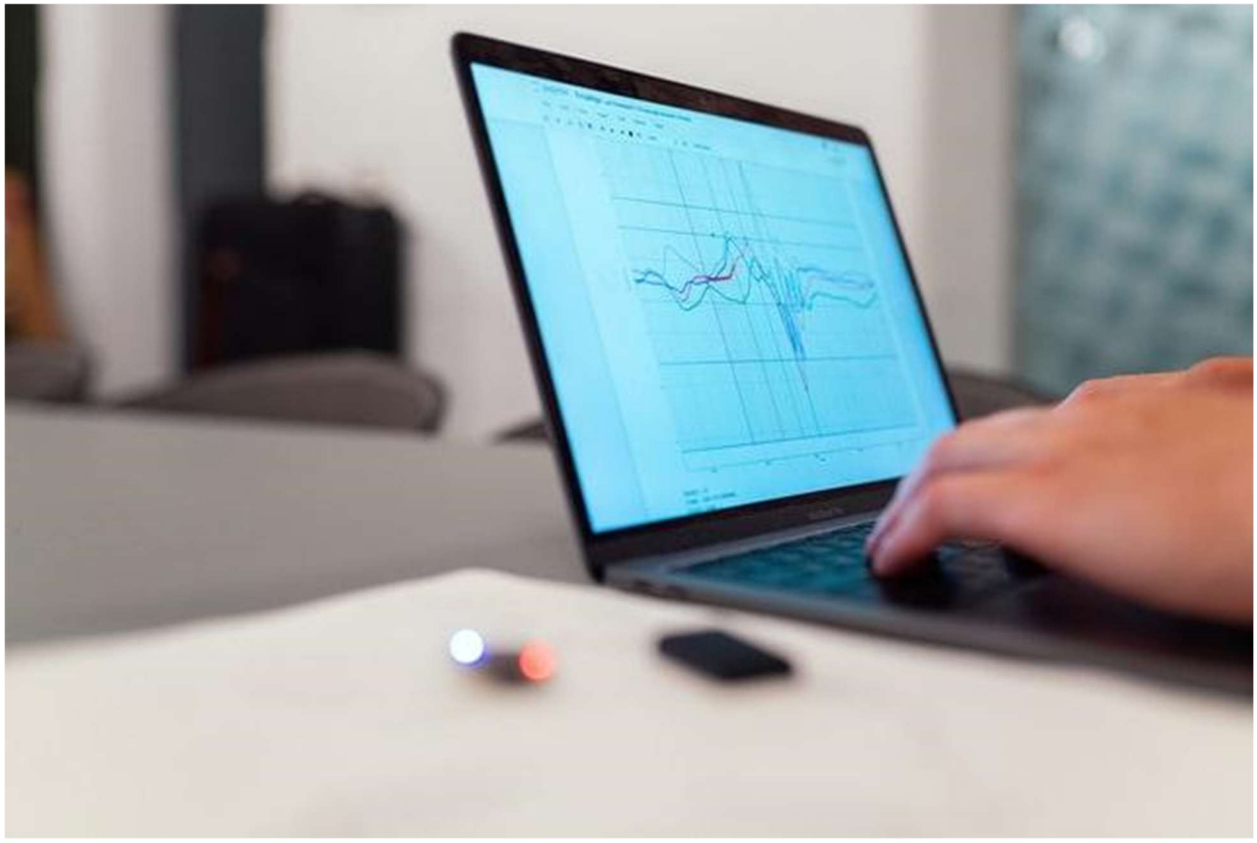


Photo by ThisisEngineering RAEng on Unsplash

Introduction

The data science and machine learning (DS/ML) field has been growing exponentially during the past 10 years. New positions have been open in basically any industry, and Data Scientists are in large demand. But not all candidates and positions are the same. The field includes experienced analysts, programmers and researchers switching into this growing field, data scientists who started in the field from the beginning and have now significant DS/ML experience, and a large number of young people just coming out of school. At the same time, the DS/ML field is composed of a large variety of tools developed in different languages, with language specific packages used in different environments, allowing to process and analyze any data from any field. As a result, skills needed to succeed in this field can range significantly. As a reflection of this large variety of skills demanded we find a wide range of salaries being offered. Job titles, though highly correlated with earnings show a large overlap. So, if you are going to hire a new recruit, how much should you offer them? We can also ask, if you want to enter the Data Science job market, what skills or characteristics will help you land those high paying jobs?

The BLS reports an average salary for Data Scientists and Mathematical Science Occupations (SOC code 152098) in the US of $103, 930, and a median salary of $98,230 with an inter-quantile range of $32,140. Other related positions add to this variation with $86,200 median salary for Operations Research Analysts (SOC code 152031) at the bottom, and on the other end, $110,140 median average for Software Developers and Software Quality Assurance Analysts and Testers (SOC code 151256).

We can find market description analysis based on job posting analysis, using actual but mainly estimated salary offers, and applying NLP tools to identify skills required for the positions (see Predict Data Science Salaries with Data Science by Junting Lai). There are few data sources that describe current positions and salaries using small surveys, with obscure methodologies (see The Burtch Works Study, Salaries of Data Scientist & Predictive Analytics Professionals, August 2020). We use a relatively large survey specifically targeting the data science and machine learning community conducted by Kaggle, and aimed to get a comprehensive view of the state of the data science and machine learning field all around the world. The survey includes a question on yearly compensation, together with respondent's demographic characteristics, experience and skills.

Before we look at the data – Where is the data coming from?

Kaggle launches a Data Science and Machine Learning (DS & ML) Survey every year as a way to learn about the DS field, with questions on the user demographics (age, gender, education level, working status), their data science and machine learning knowledge and experience, and methods and tools they use or would like to get familiar with. The survey was distributed to the entire Kaggle community through the Kaggle (opted-in) email list, and promoted on the Kaggle website and Kaggle Twitter channel.

There are 20,036 answers to the survey in 2020 from 171 different countries, including active workers, unemployed individuals and students. The number of active workers in the survey is 13,213, and after dropping observations with missing income, the sample size is reduced to 10,729 individuals.

We identify a subset of survey questions that could be relevant to predict salaries, and we transform survey answers so they will be easier to process. Multiple answer questions are transformed into a set of indicator variables for each of the possible answers. Answers that are categorical, but numeric in nature (e.g. age group) are transformed into continuous variables using range midpoints.

Observations with missing values are dropped from the analysis, with the exception of the use of ML/DS tools, with a list to choose from (one or many). Each question has an option for "Other" or "None", but given the high number of non-responses in these questions, we assume that a none response implies "None".

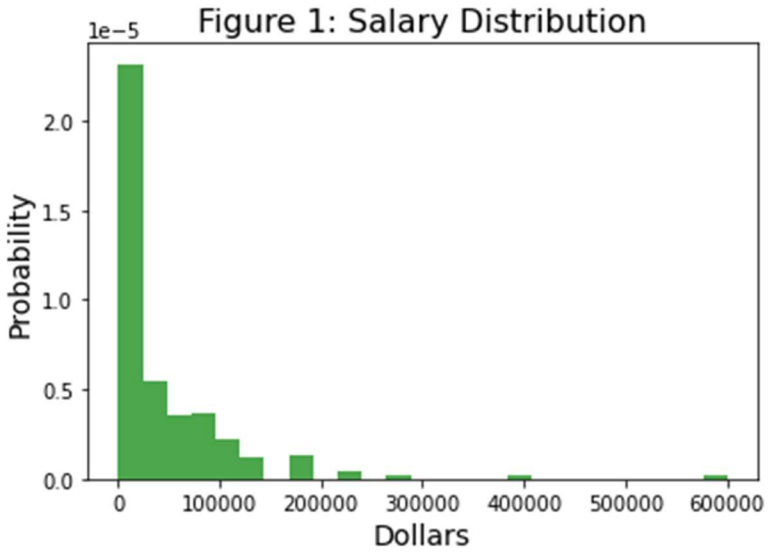
Our target variable is defined with the question: "What is your current compensation (approximate $USD)?" Note that here we use compensation and salaries indistinctly. In order to model salaries we want to consider five types of variables: demographic characteristics, geographic region, job position, experience, and DS/ML tools. Demographic variables considered are age level, education, and gender. To account for salary disparities around the globe, we created indicators for top 10

respondent countries (India, US, Brazil, Japan, Russia, UK, Germany, Nigeria, Spain, and Canada), and indicators for low income, medium-low income, medium-high income, and high income countries using data from the World Population Review. We identified two variables that indicate the type of position and employer for the individual: the job title, and the number of employees at the company. Variables considered to reflect experience are years of experience programming, years of experience using machine learning methods (ranges converted to continuous variables using range midpoint). We also consider the number of languages the person uses to program, indicators for specific languages (Python, R, SQL, C, C++, Java, Javascript, Julia, Swift, Bash, Matlab), the use of visualization methods, cloud computing platform, deep learning workstation or a personal computer/laptop, the use of advance statistical software (SPSS, SAS, etc.), basic statistical software (Excel, Google Sheets, etc), business intelligence software, or local development environments.

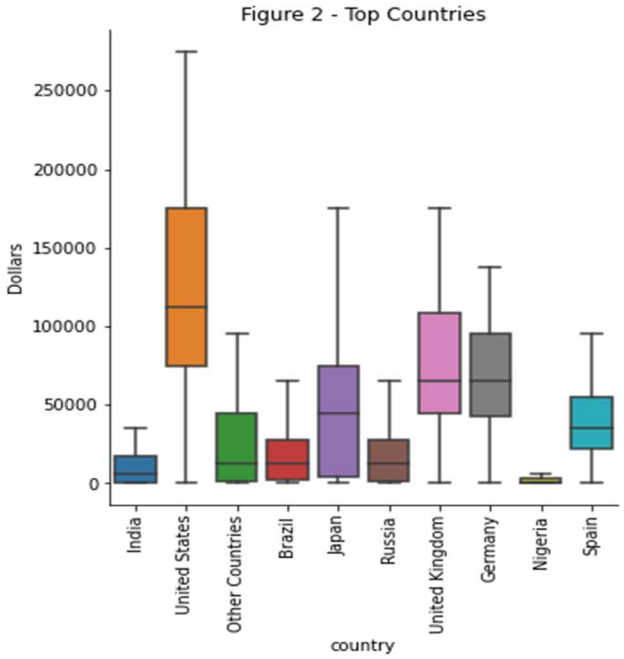
Household median income by country was downloaded from the World Population Review website, and countries were classified into four income levels using household median income quartiles.

Salary Distribution

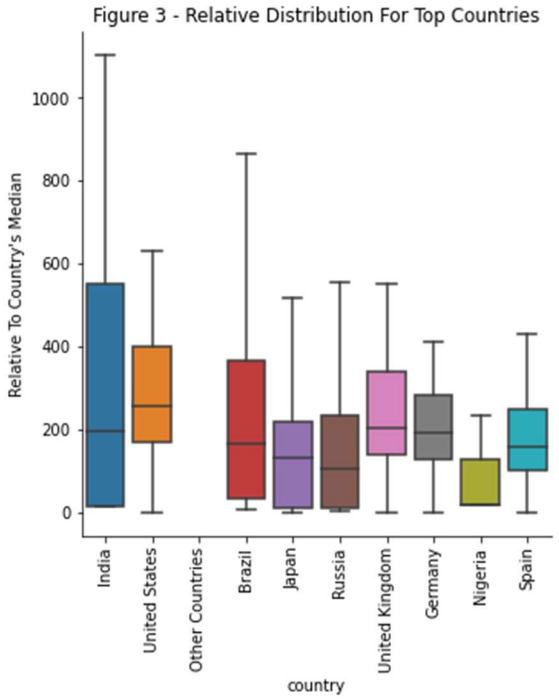
We can use the Kaggle survey to assess salaries paid to individuals in the DS/ML community. A first look at salary distribution shows large variations with a high concentration on the low end of the range, with very low compensation values. About 41% of the sample earns less than $10,000, while about 9% of the sample earns more than $100,000.



However, this plot hides large income inequalities across countries. It is now very easy to work from any part of the world for anybody else in the world. However, there are still regulations that make the substitution of workers with foreign residents harder. Figure 2 shows boxplots for salary distribution for the top ten respondent countries. Salaries in the US are much higher than in the rest of the countries, including other rich countries like the United Kingdom and Germany. US salaries also present larger variability.



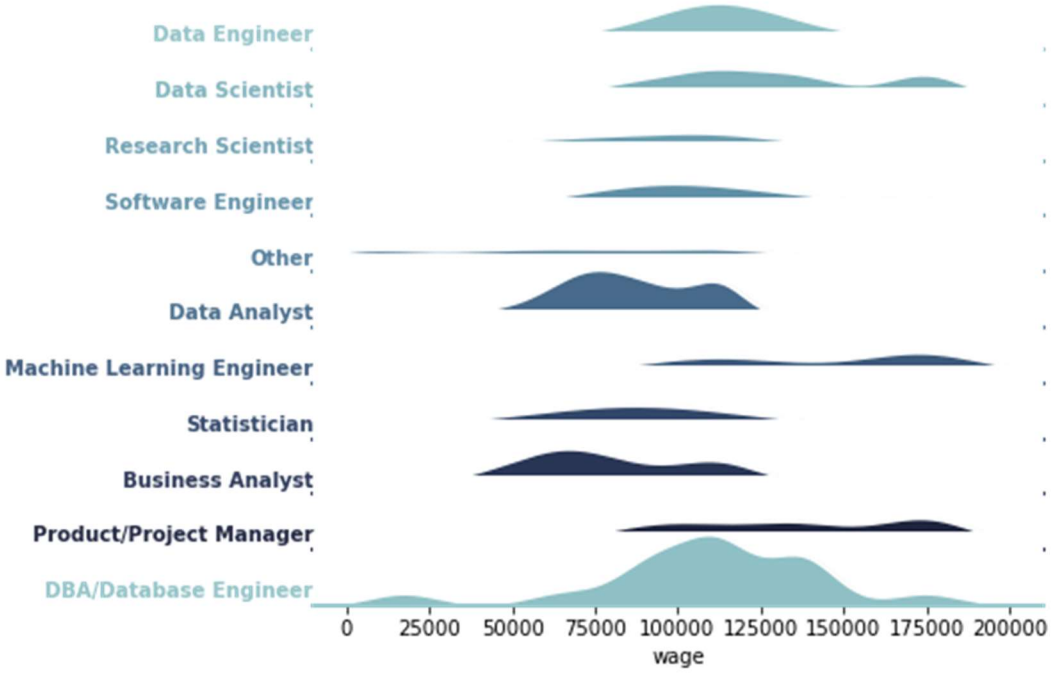
A more insightful approach when comparing salaries in the data science community across countries, is to look at wages relative to countries median income. Household median income by country was downloaded from the World Population Review website. Figure 3 shows that with the exception of Nigeria, median salaries for the DS/ML community are above countries' median income levels, with the United States, United Kingdom and Germany being more than double their median income.



The two countries with the most responses by far are India with 2,353 responses and the US with 1,484 responses. These are also the two countries on the extreme of the compensation levels, with the US having an average salary of $121,130, six times larger than India's $18,503 average salary, and a median salary of $112,500, eighteen times larger than India's $6,250 median salary.

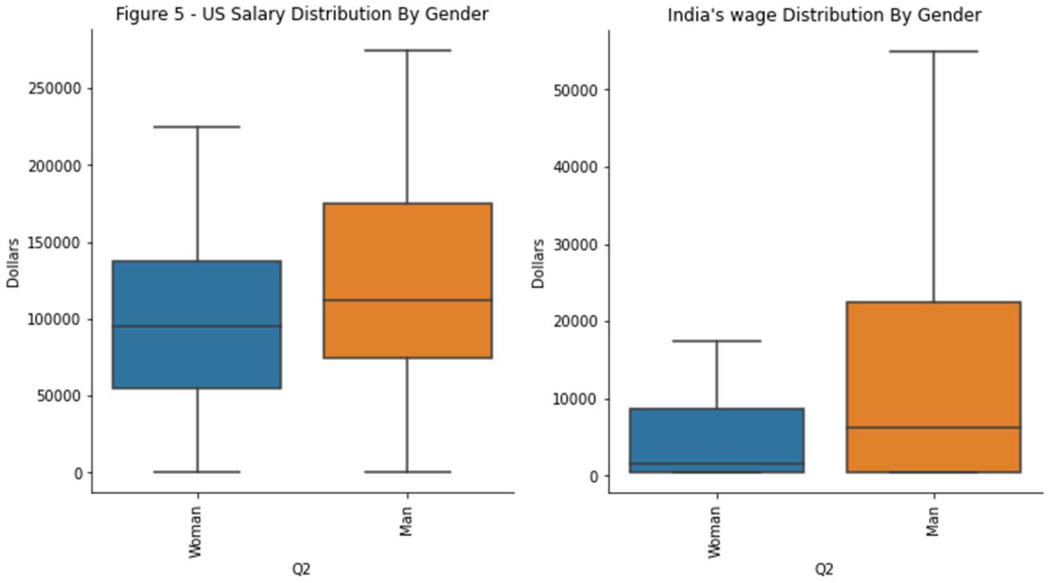
Salary distribution for a given job title, even within a given country, has a large variation. Different positions can be associated with higher or lower salaries, but there is still large overlap between job titles. Look at figure 4, with salary distribution in the US. Database administrators and engineer compensation expands over a wide range, going from a low mode below the bulk of all the other job positions, to a high mode on top of all other positions. The three other titles that present a mode on the higher end of the distribution (around $175K) are Data Scientist, Machine Learning Engineer, and Product or Project Manager. Job titles with most of their compensation distribution being on the lower end of the range are Data Analysts, Business Analysts and Statisticians.

Figure 4 US Salary Distribution By Job Title

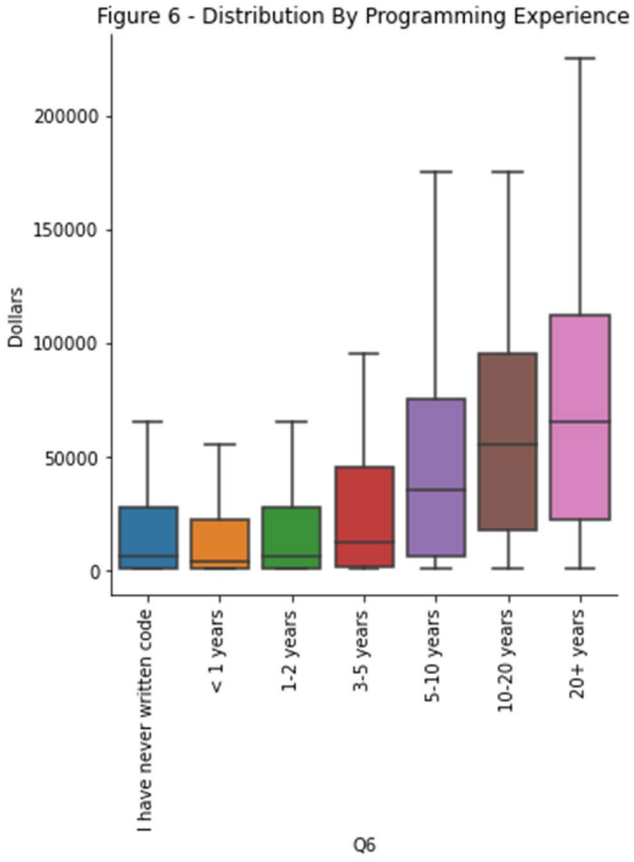


From the distribution chart we can also observe that the most frequent position observed is the DBA/Database Engineer, followed by Data Analysts, and the multiple modes observed are probably related to the country of residence.

Another common source of salary differences is gender. Gender disparity is clearly present in the DS/ML community. The salary gap between men and women is strong in this field, with most survey participants being males with median male compensation of $22,500, more than three times larger than women, and average compensation of $46,771 against $33,340 average compensation for women.



The gap is also present when we look at the two examples, India and the US, being larger in India than in the US. Still, US men salaries in the DS/ML community are on average 32% higher than those for women, with the top quartile of men reaching salaries around $175K or above, while women's top quartile earnings are around $137.5K or above.



Finally, we will expect that more experienced programmers will earn higher salaries. This is true for our sample and can be observed in figure 6, but the boxplots also show an increasing variability as experience grows. These large differences between years of experience are also present when we look at just the US. An individual with just one or two years of experience earns in the US a median salary of $75,000, while somebody with 10 or more years of experience earns a median salary of $137,500, or on average $73,430 more.

Modeling Salaries

We try two different ML modeling approaches to predict salaries, linear models and random forest. The survey is very extensive and we have a long list of potential features that can be used to predict salaries, but this can result in model overfitting. To mitigate this problem we will use cross validation methods to evaluate model fit and select the best method to predict salaries, and we try two regularization models, the Ridge Regression and the LASSO Regression. For the random forest fit we first do a grid search to optimize the parameters before modeling salaries. Given the high level of skewness in the distribution of the outcome variable we transform it using natural logs. We also normalize the matrix of covariates using the Scikit Learn scaler.

We primarily use the coefficient of determination (R2), the mean squared error, and mean absolute error (MAE) to evaluate and compare the models. A lower MSE and MAE implies a better fit. The coefficient of determination, or R2 is also commonly used, it's based on squared errors and it can be interpreted as a proportion. The R2 compares the model fit with a prediction based on simple averages. A higher R2 implies a better fit. These three measures are part of the scikit-learn library.

We first fit two linear models, a basic model with just demographic variables, experience, and country of origin, and a full model where we also include information on the type of position and tools used by the individual. This can help in evaluating how much the additional features add to the predictive power of the model. We fit the regular linear regression model, and also the Ridge and Lasso regression models that can help the fit when the covariates are correlated. The model significantly improves when using the full set of covariates, with a shift from 1.29 to 1.21 in the MAE and from 0.36 to 0.42 in the R2 for the regular linear regression. The Ridge or Lasso regression don’t improve the fit, so we run the regular regression on the training dataset and get a final evaluation using the test data. The R2 for the linear fit is 0.42 when evaluated on the test set.

The linear fit may not be the best approach given the non-linear nature of the salary and covariates relationship. We can model and predict salaries using the Random Forest, a type of bagging method that relies on the strength of fitting multiple trees and averages. The random forest can randomize not only samples but also the features used to build each of the trees.

We first fit the model with some basic and default parameters, and find that a random forest regression model improves the fit when compared with the linear model. But before looking further, let's see if the model can be improved by running a grid search to identify the best parameters for the forest. We try an extensive search over all parameters in the random forest model. We first create a grid with points for each parameter, to fit the random forest and then we use the randomized search (RandomizedSearchCV) estimator, that will pick a random set of parameters

from the grid and fit the model applying the cross validation method. Once the function finds the optimum set of parameters, we run a comprehensive search around them to refine the optimum point.

	Model	MSE	MAE	R2
0	Linear Regression	2.304	1.201	0.421
1	Ridge Regression	2.304	1.201	0.421
2	Lasso Regression	3.979	1.728	-0.000
0	Random Forest - Basic	2.138	1.159	0.436
1	Random Forest - Optimized	2.129	1.148	0.465

The optimized forest fit with the identified parameters from the grid search shows a large improvement in the model fit. Using the R2 measure it's a clear improvement not only on the linear model fit, but also on the Random Forest fit using the default parameters. With the final identified parameters we train the Random Forest and evaluate it on the test set. The table comparing the results is created using the full sample, split in two datasets, a training set and a test set. MSE, MAE and R2 are evaluated on the test dataset. We were able to decrease the MSE by 7% and increase the R2 by almost 10%.

The Results

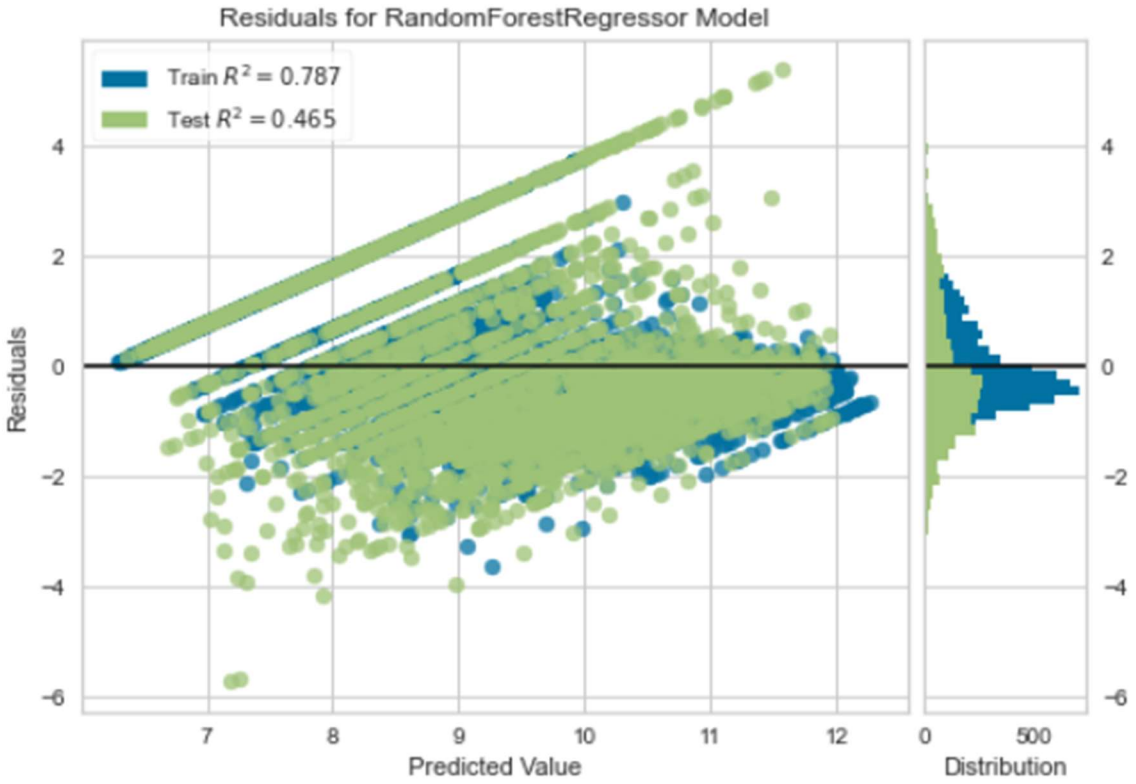
The final model gives us insight on what characteristics are more relevant when trying to predict a person's salary. We plot the top 20 features identified in this final model.

The Random Forest model using all covariates has the best fit. There is still a large amount of unexplained variation, with a significant error, not explained by the covariates. However, there are a few features that stand out in their explanation power. Of course, the country of residence is highly

relevant. In particular, being from a high-income country, or a US or Nigeria residence indicators, are among the top explanatory variables. The second set of variables with high prediction power are related to experience: age, programming experience, significant experience (more than 2 years) in machine learning methods. Employer size is another highly relevant variable. Finally, the knowledge and use of some specific tools or environment have also significant predictive power.

But we can ask: how much is the error in terms of dollars? We can compute the error after transforming the salaries back to dollars, and compute the MAE, which gives a measure in dollars. This error is on average $25,000, which seems reasonable for large salaries but too large for salaries on the lower end of the distribution.

.



The errors from the fit are clearly not random and still quite large. The Random Forest still seems to overestimate salaries when they are on the lower end of the spectrum, and underestimate them when they are on the high end of the spectrum.

This dataset has information not present in other datasets, but there are important aspects of the hires not reflected in the data. The industry is an important factor, not considered here. Some industries are well known for paying higher salaries than others. The high tech industry traditionally pays higher salaries, but the industry has been penetrating other sectors, with those jobs not always catching up to the core of the industry. This is probably the case with DS/ML positions, so adding industry information to the survey will be useful. The second big piece of information missing here is the job status. The survey asks about current job title, but there may be some people working part time or under contracts, with salaries varying accordingly.

Conclusion

The DS/ML jobs are paid in general above each country's median income, but they present large variation depending on the country of residence. Even after correcting by country, and country median income, there is still a large amount of variability in salaries. The current model uses the DS/MS Kaggle survey for 2021 which has rich information on individual's demographic and use of DS/ML tools. Applying ML algorithms to the data allows to identify top features important in predicting salaries However, the survey still falls short to explain the large amount of variation in salaries. The Kaggle survey is distributed to the community every year. We suggest the addition of two questions to the survey, which will generate highly useful information with little cost: job status (full time, part time, freelance, etc) and work industry.

Data Sources

- Kaggle 2021 Data Science and Machine Learning Survey: https://www.kaggle.com/kaggle/kaggle-survey-2018
- 2021 World Population Review: worldpopulationreview.com
- BLS, Occupational Employment and Wage Statistics https://data.bls.gov/oes/#/geoOcc/Multiple%20occupations%20for%20one%20geographical%20area