Shivangi Mundhra
SUID - 842548148

Full Time USA Data Science Job Salary Prediction

IST 716 – Applied Machine Learning | Fall 2022 | Shivangi Mundhra | SUID – 842548148

Introduction

As a master's student of Data Science graduating in less than 5 months, I cannot wait to get started with my career in Data Science and put all acquired knowledge to good use. While fantasizing about my career and job after school, the subject of pay always comes up. I am sure there are thousands of prospective graduates like me thinking of prospective pay after graduation. Whether a professional in data science is changing jobs for professional growth or other reasons, job's salary is always a major deciding factor.

As with jobs in any other domain, the salaries for data science jobs depend on a lot of factors. There's a lot of variation in the amount you can expect to make depending on your major, gender, age, education level, experience, location of employment, location of company, type of job and more. However, realizing the baseline for pay for a job at any point will prepare us for the upcoming lifestyle and changes. For instance, understanding the baseline pay will help in planning for key expenses like apartment rent and food. We want people starting new jobs to make better informed decisions. Therefore, it becomes important to know this baseline pay.

Objective

The objective of this project is to predict base salary (USD) for US full-time data science jobs depending on input factors of

1. company size
2. job category, and,
3. experience

Through this report, I am going to walk through all the aspects of this project including the problem at hand, the data source and the data set, the descriptive analysis and creating and testing different prediction models. In the analysis itself, I will try to explain all reasoning and thought process behind each significant step. The intent is that any person, especially one without technical knowledge and that is not familiar with this problem statement, should be able to read this report and follow along.

The salary variable is of continuous type. In order to achieve our goal, we will have to use regression analysis. Therefore, we will use different regressors and evaluate their performance to find the best model. The target variable should be salary (USD) and the predictors will be company size, job category and experience.

Shivangi Mundhra
SUID - 842548148

Evaluation Techniques

We will create different Machine Learning models to predict the salary depending on different factors. There are a lot of metrics that ML models could be evaluated over. However, since our models will use regression, we will evaluate these models on the below metrics

Mean Absolute Error

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

Root Mean Squared Error

RMSE is the square root of the average of squared differences between prediction and actual observation. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. RMSE increases with the variance of the frequency distribution of error magnitudes.

A combination of these two metrics for each model should present us with a clear view of how well the model is performing.

The Data Source

The data for this project was obtained from ai-jobs.net/salaries. This site collects salary information anonymously from professionals all over the world in the AI/ML/Data Science space and makes it publicly available for anyone to use, share and play around with. The website's primary goal is to have data that can provide better guidance regarding what's being paid globally. So, newbies, experienced pros, hiring managers, recruiters and startup founders or people wanting to make a career switch can make better informed decisions.

This dataset is published in the public domain under CC0, which means that we can use, copy, modify and distribute the work without asking for permission. The raw data downloaded on 12/04/2022 contains of 1,637 records of data. Below is the column data dictionary. We will not use all columns for our models and not all records. Since we are only looking for predicting full-time salaries for US based employees, we will filter on employment_type = 'FT' and employee_residence = 'US'. After applying these filters, we receive 1,167 usable records.

Column Name	Description	Data Type	Example Values
work_year	The year the salary was paid	Integer	2020, 2021, 2022
experience_level	The experience level in the job during the year	String	EN - Entry-level / Junior MI - Mid-level / Intermediate SE - Senior-level / Expert EX - Executive-level / Director
employment_type	The type of employement for the role	String	PT - Part-time, FT - Full-time, etc
job_title	The role worked in during the year	String	Data Analyst, Data Scientist, Machine Learning Engineer, etc
salary	The total gross salary amount paid in local currency	Integer	
salary_currency	The currency of the salary paid as an ISO 4217 currency code	String	USD, EUR, GBP, etc
salary_in_usd	The salary in USD	Integer	
employee_residence	Employee's primary country of residence in during the work year	String	US, IN, CA, UK, etc
remote_ratio	The overall amount of work done remotely	Integer	0 - No remote work (less than 20%) 50 - Partially remote 100 - Fully remote (more than 80%)
company_location	The country of the employer's main office or contracting branch	String	US, IN, CA, UK, etc
company_size	The average number of people that worked for the company during the year	String	S - less than 50 employees (small) M - 50 to 250 employees (medium) L - more than 250 employees (large)

Shivangi Mundhra
SUID - 842548148

Data Pre-processing

As mentioned earlier, after filtering the data on employment_type = 'FT' and employee_residence = 'US', we received a total of 1,167 usable records for our analysis. Since we are not using the work_year as a predictor, we need to convert the USD salaries of all our prior work_year (2020 and 2021) records to a level of 2022 work_year. To obtain this, we take inflation into account. According to the US Bureau of Labor Statistics and data obtained from in2013dollars.com, the average inflation rate was 4.7% per year between 2020 and 2021, and, 7.75% per year between 2021 and 2022. To obtain the equivalent 2022 salaries, we multiply the prior salaries with respective rates.

There are records of 53 unique job titles in these 1,167 full-time US based salary records. For better modeling, we categorize these 53 job titles in 4 broad job categories –

1. Analyst – includes jobs like BI Analyst, Business Data Analyst, Data Analytics Consultant, Data Specialist, Machine Learning Developer and more.
2. Scientist – includes jobs like Applied Data Scientist, 3D Computer Vision Researcher, AI Scientist, Lead Data Scientist and more.
3. Manager – includes jobs like Data Analytics Manager, Data Science Manager, Data Manager, Head of Data, Head of Data Science and more.
4. Engineer – includes jobs like Data Engineer, Machine Learning Engineer, Data Architect, ML Engineer, Data Operations Engineer and more.

We remove the columns that are not required for this project. The final data frame only contain the below columns

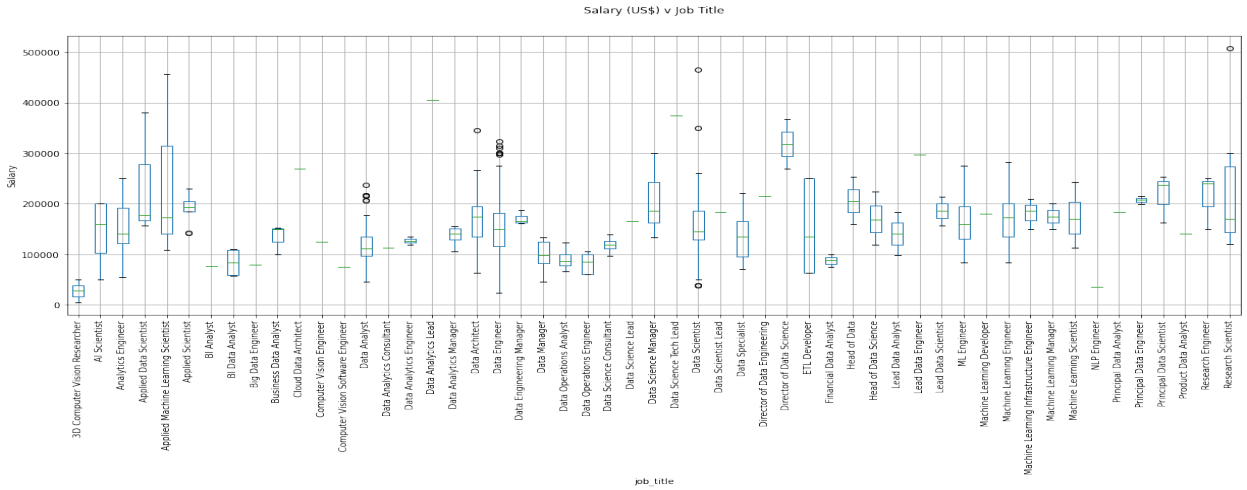
1. Experience level
2. Job Categories
3. Company Size
4. Salary

In order to train and test our model, we split our data in training and testing data frames in 80:20 ratio. We use 933 records to train models and 234 records to test. Salary is the target variable and rest are predictors.

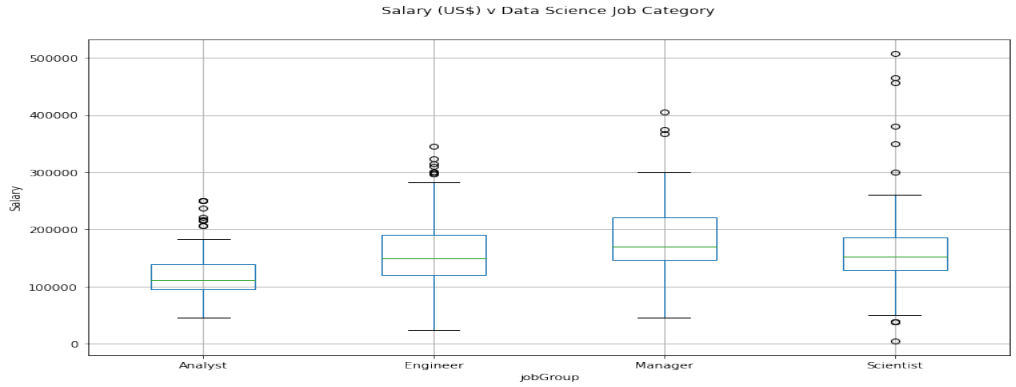
Descriptive Analysis

The below chart shows the boxplots of salaries in USD for all full-time US based jobs. As we can see, the largest salary range corresponds to the job title of Applied Machine Learning Scientist, the highest paid job is for a Research Scientist and the lowest salary is for a 3D computer vision researcher.

Shivangi Mundhra
SUID - 842548148

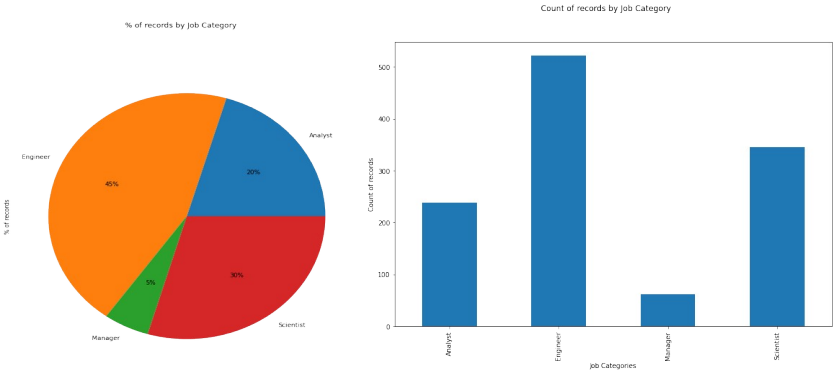


After converting these different job titles to job categories, we plot the salaries boxplots again, this time against the job categories and we obtain the below.



As we see, the job group of Scientist has the highest paid salary, however, the Inter-Quartile Range (IQR), the range of salary between 25 percentile of jobs and 75 percentile of jobs is at the highest for Manager job group. The Manager job category also has the highest median salary.

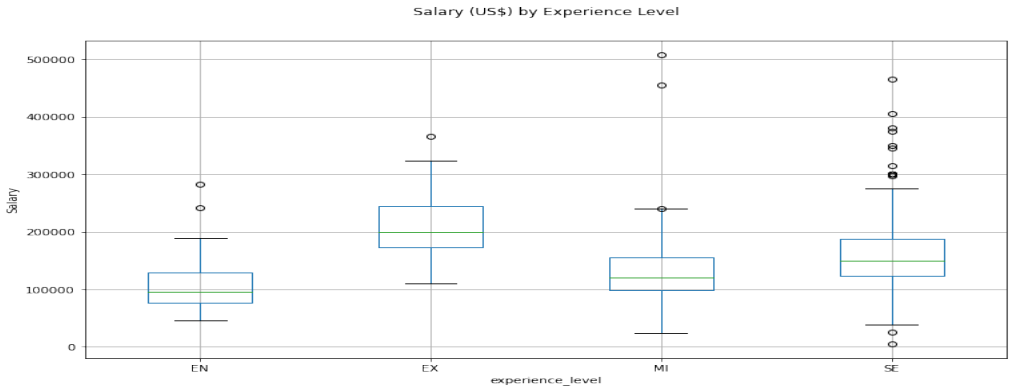
The below charts show the distribution of count of records by job category.



Shivangi Mundhra
SUID - 842548148

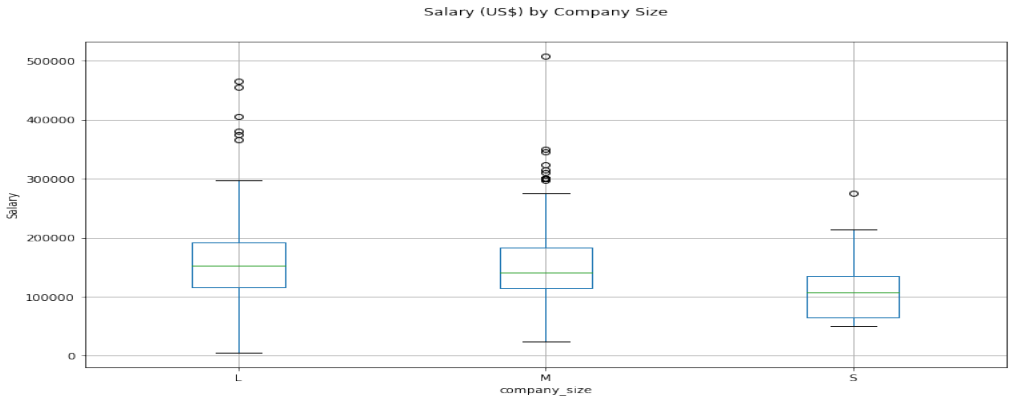
Engineer job category has the largest number of records with over 500 records in that category. Manager jobs comprise of only about 5% of all records.

Below chart shows boxplots of salaries in USD by experience level.



Predictably, entry level box has the lowest box followed by mid-level box followed by senior and the box for executive salaries is the highest. The median salaries also follow the same order. However, the highest salary is attributed to a mid-level record. This could be an outlier because of a few people getting paid more in some companies due to niche skills.

Below chart shows the boxplot of salaries in USD by company size.



Again, this follows a predictable behavior. The box for small companies is the lowest and the box for large companies is the highest indicating that, in general, large companies pay the most and small companies pay the least.

Prediction Models

We created the below machine learning models and trained them using the training dataset. For tuning the model, we used grid search method to find the parameters which give us the best model.

Shivangi Mundhra
SUID - 842548148

Linear Regression

In this model, the target value is described in an equation in terms of predictors. An intercept and coefficients are calculated for each predictor using the training data. The predictions are then made for test dataset using this equation and plugging in the values of predictors in the equation. These types of models are extremely easy to interpret but have limitations because not all real data can be described with linear trend.

Below are some parameters that were included in tuning this model.

1. Fit_intercept - controls whether the model should create an intercept. Default = True.
2. Positive – if this parameter is set to true then all coefficients will be forced to be positive. Default = False.
3. N_jobs – determines the number of jobs to use for computation for speeding up the prediction. Default = None.

Decision Tree Regressor

In this model, a tree of decisions is created based on values of predictors. The model predicts the value of target variable using these steps of decisions. These models are easy to interpret as well and can be visualized in terms of a tree diagram.

Below are the parameters that were included for tuning.

1. Max_depth – the maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. Default = None.
2. Min_samples_leaf – the minimum number of samples required to be at a leaf node. Default = 1.
3. Splitter – ‘best’ splitter chooses the best split at each node and ‘random’ splitter chooses randomly. Default – ‘best’.
4. Max_features – the number of features to consider when looking for the best split. Default = ‘auto’ selects max_features = number of features, ‘sqrt’ selects max_features = sqrt(n_features) and ‘log2’ selects max_features = log2(n_features).

Random Forest Regressor

A random forest is a meta estimator that fits several classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Below are the parameters that were included for tuning.

1. N_estimators – decides the number of trees in the forest. Default = 100.
2. Max_depth – the maximum depth of the tree. If None (default), then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
3. Criterion – the function to measure the quality of a split. Supported criteria are “squared_error” for the mean squared error, which is equal to variance reduction as

feature selection criterion, "absolute_error" for the mean absolute error, and "poisson" which uses reduction in Poisson deviance to find splits.

K Nearest Neighbors Regressor

In this model, the target is predicted by local interpolation of the targets associated of the nearest neighbors in the training set.

Below are the parameters that were included for tuning.

1. N_neighbors – number of neighbors to use. Default = 5.
2. Weights – the weight function used in prediction. 'uniform' : uniform weights. All points in each neighborhood are weighted equally. 'distance' : weight points by the inverse of their distance. in this case, closer neighbors of a query point will have a greater influence than neighbors which are further away.
3. Algorithm – algorithm used to compute the nearest neighbors. Default = 'auto' will attempt to decide the most appropriate algorithm based on the values passed to fit method. 'brute' will use a brute-force search.

Model Comparison and Evaluation

The below table shows the performance in terms of errors for each prediction model. The tuned model's parameters are listed in the 'Parameters' column. Only those parameters are listed that vary from the default model. Other parameters were included in hyperparameter tuning of each model, however, since, the parameter tuning did not affect some parameters, we have not included those in this table.

Rank	Model	Parameters	Root Mean Squared Error	Mean Absolute Error
1	Decision Tree Regressor	Tuned - Max_depth = 5 min_samples_leaf = 2	$ 49,160	$ 36,967
2	Random Forest Regressor	Default	$ 49,393	$ 37,485
3	Random Forest Regressor	Tuned - n_estimators = 25	$ 49,704	$ 37,609
4	Decision Tree Regressor	Default	$ 49,549	$ 37,631
5	kNN Regressor	Tuned - algorithm = 'brute' n_neighbors = 72 weights = 'distance'	$ 49,656	$ 37,848
6	Linear Regressor	Default	$ 52,030	$ 40,157
6	Linear Regressor	Tuned - Same parameters as default model	$ 52,030	$ 40,157
8	kNN Regressor	Default	$ 57,532	$ 45,181

The best model is achieved using hyper-parameter tuning of decision tree regressor. The tuned model has max_depth = 5 instead of the default value of 'none' and a min_samples_leaf = 2 instead of the default value of 1.

Shivangi Mundhra
SUID - 842548148

Understanding the Implications of Error

The mean absolute error of the most efficient model is $36,967. The same model's root mean squared error is $49,160. This means that the prediction from the model could be off by up to ~$37k. The main cause of this error is clearly the high variance of salaries in each type. For instance, according to the data, a mid-level (experience) data analyst (job title) at a medium size company can earn anywhere between $50,000 and $216,000 in a year with 50% of records falling between $100,000 ($25^{th}$ percentile) to $150,000 ($75^{th}$ percentile) and a median of $124,000. Let's try to make sense of this given our filters on the dataset.

We have used only US salary data. This means that we are categorizing an employee residing in Syracuse working for a medium sized company as a mid-level data analyst in the same category as an employee residing in the San Francisco Bay Area working for the same medium size company at the same experience level. It is highly likely that the salary of the employee residing in the San Francisco Bay Area is greater than the salary of employee residing in Syracuse by $35,000 - $50,000. Therefore, the error values in our tuned decision tree prediction model make sense.

One way to get a more accurate prediction model would be to get a data source that also indicates the US state of employee residence. This would significantly decrease the variance in salaries and help decrease MAE and RMSE.

Conclusion

Based on the analysis and findings in this report, we can conclude the following –

1. We can predict the salary for different full-time data science jobs in the USA using machine learning models. The performance of these models depends on the type of model, and, to some extent, on the parameters that we use to define these models. In our case, the best results were obtained using a tuned decision tree model. This is good because decision tree models are easily interpretable and do not require explaining the black box machine learning models like for Support Vector Machines or Neural networks.
2. We do receive an error of $36k which can be attributed to the variance in the same job salaries in the same sized company with the same experience level. As from our explanation above, this kind of variance can be avoided if we add state/city/territory information as predictors.
3. It is important to report the error when reporting the prediction for a set of input variables. One can use the error to make better informed decisions.

References

1. ai-jobs.net/salaries/download/
2. US Bureau of Labor Statistics
3. In2013dollars.com
4. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html