7/31/2022

Tran Minh Khoa

2101432I

MLDP Project

Report

# Introduction

## Topic: Data Science Job Salaries

Jobs in the field of Data Science vary widely in salary. This project aims to determine how various factors – such as job title, company size, and country of work – impact annual salary.

## Dataset

**URL:** <https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries>

**Shape:** 607 rows, 12 columns

**Columns:**

|  |  |  |
| --- | --- | --- |
| Column | Type | Description |
| company\_location | Nominal | Country Code1 of company’s contracting branch (50 unique values) |
| company\_size | Ordinal | Number of people that worked for the company  *With being the number of employees,*   |  |  | | --- | --- | | Value | Meaning | | S |  | | M |  | | L |  | |
| employee\_residence | Nominal | Country Code1 of employee’s primary residence (57 unique values) |
| experience\_level | Ordinal | Level of experience holding *job\_title*   |  |  | | --- | --- | | Value | Meaning | | EN | Entry-level / Junior | | MI | Mid-level / Intermediate | | SE | Senior-level / Expert | | EX | Executive-level / Director | |
| job\_title | Nominal | Role of work (50 unique values) |
| remote\_ratio | Ordinal | Percentage of work done remotely  *With being the percentage of remote work,*   |  |  | | --- | --- | | Value | Meaning | | 0 |  | | 50 |  | | 100 |  | |
| salary\_currency | Nominal | Currency of *salary* |
| salary | Interval | The salary paid |
| salary\_in\_usd | Interval | *salary* converted to USD |
| work\_year | Interval | Year when salary was paid |

1 [ISO 3166](https://www.iso.org/iso-3166-country-codes.html)

2 [ISO 4217](https://www.iso.org/iso-4217-currency-codes.html)

# Data Exploration & Pre-processing

## Preprocessing

### Cleaning

|  |  |  |  |
| --- | --- | --- | --- |
| Issue | Description | Fix(es) | |
| Duplicates | There were no duplicate indexes, but 71 rows with duplicate values.  These duplicates were sparsely distributed (usually 2 per unique duplicate). | *Ignore*  They probably refer to employees working for the same company, in the same role. | |
| salary & `salary\_currency` | These are used to calculate, and hence directly related to, `salary\_in\_usd`. | *Drop*  `salary\_in\_usd` can accurately represent them. | |
| Low Variance columns | `employment\_type` is dominated by “FT” (96.9% of values) | *Drop* | |
| Missing values | There were no missing values | | |
| Outliers in `salary\_in\_usd` | 16 values were over 2 standard deviations away from the mean | | *Drop*  Outliers sway modeling |

### Transformation

The general motive is to reduce noise & numericize for better [Modeling](#_Methods_&_Improvements).

|  |  |
| --- | --- |
| Transformation | Justification |
| *Integer-ize* `company\_size` | Ordinal columns can be Numericized as rank numbers |
| *Integer-ize* `experience\_level` |
| *Boolean-ize* `employee\_residence`  *Rename* as *`ricl`* (*r*esides *i*n *c*ompany *l*ocation) | Has many similarities with `company\_location`; it would be more helpful to know, not where the employee lives, but whether they live in the same country as their company. |
| *Bin* `company\_location` by continent | Noise Reduction of 50 unique values to 6 |
| *Integer-ize* `company\_location` from West to East   * The Westmost, NA, is 1 * The Eastmost, OC, is 6 | Attempt at Numericizing |
| *Standardize* `job\_title` with the following [RegEx](https://en.wikipedia.org/wiki/Regular_expression):   1. /.\*Researcher/ ⇒ “Research Scientist” 2. /AI|Computer Vision|Machine Learning|NLP/ ⇒ “ML” 3. /.\*(Data|ML).\*(Analyst|Architect|Engineer|Researcher|Scientist).\*/ ⇒ “$1 $2” 4. /(Director|Head).\*|.\*Manager/ ⇒ “Manager”   *Bin* titles with less than 5% frequency as “Other” | Noise Reduction of 50 unique values to 6;  Many titles had similarities, only different due to prefixes/suffixes or acronyms, etc.  This approach assumes that the standardized titles are truly similar. |

|  |  |
| --- | --- |
| *Popularity Encode* `job\_title`  *Rename* as `job\_popularity` | (Singh, 2020)  Attempt at Numericizing;  Better than *Integer Encoding*, which gives false value to the difference between titles |
| *Min-Max Normalize* all numeric columns except `work\_year` | [Modeling](#_Methods_&_Improvements) will assign weights. Normalizing allows weights to be interpreted on the same scale (as significances)  Years outside this dataset should still be able to be passed into models, so Min-Max scaling would not make sense on `work\_year`. |

### Correlation Analysis

If there exists a pair of strongly correlated columns (), either column can be used to accurately estimate the other, so only one of the pair will be needed.

However, there are no columns with either strong Linear or strong Rank correlation (statslectures, 2010).

Table

Description automatically generated

## Exploration

### Most employees live in the same country as their company

A potential reason for this could be that Companies pay higher to employees living in the same country.

Chart, pie chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

### Most Data Science jobs come from North America

This majority is consistent even when divided by population[[1]](#footnote-1).

Chart, pie chart

Description automatically generated Chart, pie chart

Description automatically generated

This could match with the fact that the US consumes the most data in the world (Chakravorti et al., 2019), meaning that NA has the most resources and opportunities available for Data Science jobs.

However, this could also suggest a bias in the data. The distribution may change if more data is collected from all around the world.

### Salary increased together with work year

Chart, box and whisker chart

Description automatically generated

This increase rate is not only to counter Inflation Rate (IR)[[2]](#footnote-2); there is an effective year-on-year increase in salary; Data Science Jobs are getting paid more by the year.

|  |  |  |
| --- | --- | --- |
|  | Increase in mean salary | US IR |
| **From 2020 to 2021** | 13% | 1.4% |
| **From 2021 to 2022** | 34% | 7.0% |

# III & IV. Methods & Improvements + Results & Analysis

## Pre-Modeling

The dataset was split into the following:

### Target: salary\_in\_usd

* + Is right-skewed, but logging makes it left-skewed.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

* + Models can be trained to target both original and logged values; the better result will be selected.
  + Predictions from models using logged values will be re-exponentiated
* Test Size: 1/3

Since the [Target](#_Target:_salary_in_usd) is a continuous variable, Classifiers were not used.

The Cost Function (Evaluation Criteria) used for all models is Squared Error (SE), presented as Root Mean SE (RMSE).

All optimization algorithms used make use of Cross-Validation, so results presented should be consistent with different sets of test data.

## Exploration of Models + Backward Feature Elimination (BFE)

Sequential Feature Selection (BFE) was used to remove features to minimize SE.

* Results using between all and half of features are compared. Using less than half of features was tested to be ineffective and assumes that most features are irrelevant to the Target.
* The selected features which yielded the lowest (best) RMSE are shown.

### Starting with Ordinary Least Squares

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Did the Original or Logged Target give better results? | Best RMSE + Removed Features (if any) | Description |
| Linear Regression (LR) | Logged | 40608.50 | Fits a straight line which yields the Lowest Squared Error (Least Squares) |
|  | | | |

### Does Regularization help?

|  |  |  |  |
| --- | --- | --- | --- |
| Ridge | Logged | 40788.88 | Introduces a bias of to LR’s Least Squares. (StatQuest with Josh Starmer, 2018a)  Useful for desensitizing the best fit line to training data. In this case, Ridge performs “worse” than LR because the train & test data are similar. |
|  | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Lasso | Original | 42093.98  remote\_ratio, work\_year, & job\_popularity | Introduces a bias of to LR’s Least Squares. (StatQuest with Josh Starmer, 2018b)  Lasso performs worse than LR for the same reason as Ridge.  Ridge can only scale features according to their importance, but Lasso can also remove “useless” features. In this case, Lasso performs worse than Ridge because feature selection has already been done beforehand. |
|  | | | |

### Can Non-linear models perform better?

|  |  |  |  |
| --- | --- | --- | --- |
| Regression Tree (RT) | Logged | 43369.06  job\_popularity, work\_year, company\_size, & remote\_ratio | Partitions data based on certain rules & predicts using partition means (StatQuest with Josh Starmer, 2019b) |
|  | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Gradient Boosting Regression (GBR) | Original | 40117.74 | Chains multiple RTs to reduce error: the error of the current model is used as the target of the next. (StatQuest with Josh Starmer, 2019a) |
|  | | | |
| Support Vector Regression (SVR) | Logged | 40292.60  work\_year | Fits a hyperplane (Raj, 2020) |
|  | | | |

## Hyperparameter Tuning

The top 3 models – GBR, SVR, & LR – were shortlisted for further tuning. The optimizing algorithm used is Grid Search.

|  |  |  |
| --- | --- | --- |
| Model | Tested Parameters | Best RMSE + Parameters |
| GBR | |  |  |  | | --- | --- | --- | | Parameter | Values | Justification | | learning\_rate (Limits how much each tree contributes to the overall prediction) | 0.1, 0.01 | Prevents overfitting | | n\_estimators  (Limits number of trees) | 30-70; interval of 10 | | max\_leaf\_nodes  (Limits number of unique predictions per tree) | 4-8; interval of 1 | | tol  (Stops boosting when SE does not improve by at least this value) | 0.1-0.3; interval of 0.05 | | 40415.86   |  | | --- | | 0.1 | | 4 | | 60 | | 0.1 | |
|  | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SVR | |  |  |  | | --- | --- | --- | | Parameter | Values | Justification | | tol  (Stops boosting when SE does not improve by at least this value) | 0.12-0.15; interval of 0.005 | Prevents overfitting | | C  (Introduces a bias to SE, reducing sensitivity towards training data) | 0.6-0.9; interval of 0.05 | | epsilon  (Prevents a penalty from being applied to values with SE within this value) | 0.1-0.5;  interval of 0.1 | Prevents underfitting | | 39685.80   |  | | --- | | 0.145 | | 0.75 | | 0.2 | |
|  | | |
| LR | No parameters to optimize | |

## The Best Model

### SVR

Parameters

* C: 0.145
* epsilon: 0.2
* tol: 0.145

Trained with `work\_year` removed and [Target](#_Target:_salary_in_usd) logged

RMSE: 39685.80

### Error

Distribution of Error is normal

RMSE is satisfactory

* Less than 20% (1/6) of the range of [Target](#_Target:_salary_in_usd)
* Less than the standard deviation of [Target](#_Target:_salary_in_usd)

Chart, histogram

Description automatically generated Text

Description automatically generated with medium confidence

# Conclusion

## Summary

By knowing the company size & location, whether one lives in the country of their company, the job’s popularity, one’s experience level, and the percentage of work done remotely, the Salary for a Data Science Job **can** be accurately calculated.

Through various transformations and cleaning, and exploring different models, Support Vector Regression was found to be the best model for predicting salaries, given the available data.

## Areas of Improvement

While the model can be considered satisfactory, there are a few areas which could be done better.

### Standardization of `job\_title`

A less assumptive method could be used to group titles, such as Clustering or Analysis of Variance. Natural Language Processing could even be employed to extract meaningful differences between titles.

### Min-Max Normalization (MMN)

MMN worked well for columns with well-defined domains – company\_size, company\_location, experience\_level, & remote\_ratio. For this dataset, MMN could be applied on job\_popularity because there were limited job titles, but the data would have to be re-transformed and the model retrained in order to predict on new popularities. The same principle goes for work\_year, though it was not used in the final model. Perhaps another form of normalization would work better.

# References

Chakravorti, B., Bhalla, A., & Chaturvedi, R. S. (2019, January 24). *Which Countries Are Leading the Data Economy?* Harvard Business Review. <https://hbr.org/2019/01/which-countries-are-leading-the-data-economy>

Raj, A. (2020, October 3). *Unlocking the True Power of Support Vector Regression*. Medium. <https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0>

Singh, S. (2020, February 24). *Categorical Variable Encoding Techniques - Analytics Vidhya*. Medium. <https://medium.com/analytics-vidhya/categorical-variable-encoding-techniques-17e607fe42f9>

StatQuest with Josh Starmer. (2018a, September 24). *Regularization Part 1: Ridge (L2) Regression* [Video]. YouTube. <https://www.youtube.com/watch?v=Q81RR3yKn30>

StatQuest with Josh Starmer. (2018b, October 1). *Regularization Part 2: Lasso (L1) Regression* [Video]. YouTube. <https://www.youtube.com/watch?v=NGf0voTMlcs>

StatQuest with Josh Starmer. (2019a, March 25). *Gradient Boost Part 1 (of 4): Regression Main Ideas* [Video]. YouTube. <https://www.youtube.com/watch?v=3CC4N4z3GJc>

StatQuest with Josh Starmer. (2019b, August 20). *Regression Trees, Clearly Explained!!!* [Video]. YouTube. <https://www.youtube.com/watch?v=g9c66TUylZ4>

statslectures. (2010, October 3). *Spearman Correlation* [Video]. YouTube. <https://www.youtube.com/watch?v=YpG2MlulP_o>

1. Obtained from <http://worldometers.info/geography/7-continents> [↑](#footnote-ref-1)
2. Obtained from <https://tradingeconomics.com/united-states/inflation-cpi> [↑](#footnote-ref-2)