SENG 474 - Project Report

Data Science Job Salary Predictions

Chris Colomb (V00970873)
Maika Rabenitas (V00970890)

Contents

Introduction

Motivation

With the emergence of the AI boom, characterized by rapid advancements in artificial intelligence and machine learning technologies, the relevance and appeal of data science have grown. More individuals are recognizing the potential of data science as a career path, leading to a surge in interest within the field.

In the current job market, the demand for skilled data science professionals is evident across various industries and regions. As organizations increasingly rely on data-driven decision-making processes, the role of data scientists has become indispensable in extracting actionable insights from vast datasets. The intricacies of job salaries within the data science domain reflect a complex interplay of factors such as experience level, employment type, and geographic location.

Moreover, the advent of remote work and flexible arrangements has reshaped traditional notions of workplace dynamics, prompting a reevaluation of compensation models. Analyzing how remote work influences salary determinants offers valuable insights for both employers and employees navigating this evolving landscape.

Furthermore, disparities in compensation based on company size, job title, and employee residence underscore the multifaceted nature of salary determination in data science. Startups may offer competitive salaries to attract top talent, while larger corporations may provide additional perks and benefits. By examining these nuances, our project aims to provide a comprehensive understanding of salary dynamics within the data science industry.

Through data collection, preprocessing, visualization, and mining, our project seeks to uncover patterns and trends that shape salary structures. Ultimately, our endeavor is to contribute to greater transparency and equity in the job market, fostering an environment where both job seekers and employers can make informed decisions.

Research Questions

General Questions

- How do different factors influence job salaries?
- Are there any notable trends or patterns in salary structures?

Specific Question

- If we were to become an entry-level data engineer in 2024, how much would our salary change working in Canada compared to the US? Does remote work play a factor?

Data Structure Analysis and Visualization

Breakdown

The decided dataset has a total of 12 features. The breakdown of the features is as follows:

- `id`: A unique identifier for each row
- `work_year`: The year the salary was paid
- `experience_level`: EN - Entry Level, MI - Mid Level, SE - Senior Level, EX - Executive Level/Director
- `employment_type`: FT - Full Time, CT - Contract, FL - Freelance
- `job_title`: The role worked in during the year
- `salary`: The total gross salary amount paid
- `salary_currency`: The currency of the salary paid as an ISO 4217 currency code
- `salary_in_usd`: The salary in USD (FX rate divided by avg. USD rate for the respective year via fxdata.foorilla.com)
- `employment_residence`: Employee's primary country of residence in during the work year as an ISO 3166 country code
- `remote_ratio`: The overall amount of work done remotely, possible values are - as follows: 0 - no remote work (less than 20%), 50 - partially remote, 100 - fully remote (more than 80%)
- `company_location`: Indicates the location of the companies where the employees work. It is specified as country codes (e.g., "US" for the United States and "NG" for Nigeria)
- `company_size`: The average number of people that worked for the company during the year: S - less than 50 employees (small), M - 50 to 250 employees (medium), L - more than 250 employees (large)

For a glimpse of the `csv` file used, here are the top 5 rows:

	id	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	0	2020	MI	FT	Data Scientist	70000	EUR	79833	DE	0	DE	L
1	1	2020	SE	FT	Machine Learning Scientist	260000	USD	260000	JP	0	JP	S
2	2	2020	SE	FT	Big Data Engineer	85000	GBP	109024	GB	50	GB	M
3	3	2020	MI	FT	Product Data Analyst	20000	USD	20000	HN	0	HN	S
4	4	2020	SE	FT	Machine Learning Engineer	150000	USD	150000	US	50	US	L

Info for each data column is as follows:

```
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
```

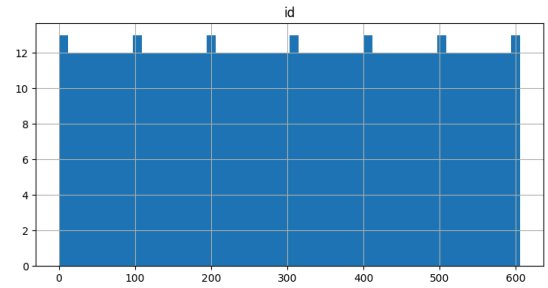
```
 0   id                  607 non-null    int64
 1   work_year           607 non-null    int64
 2   experience_level    607 non-null    object
 3   employment_type     607 non-null    object
 4   job_title           607 non-null    object
 5   salary              607 non-null    int64
 6   salary_currency     607 non-null    object
 7   salary_in_usd       607 non-null    int64
 8   employee_residence  607 non-null    object
 9   remote_ratio        607 non-null    int64
 10  company_location    607 non-null    object
 11  company_size        607 non-null    object
dtypes: int64(5), object(7)
```

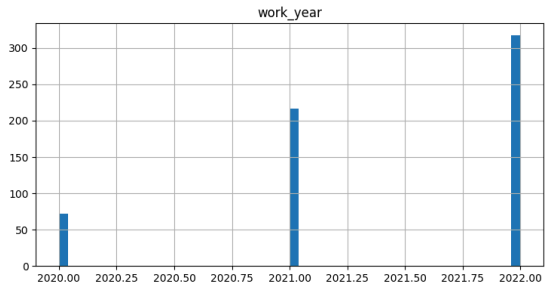
There are 607 instances in the dataset. We can see that each attribute has 607 non-null values, which means that there are no missing values in the dataset.

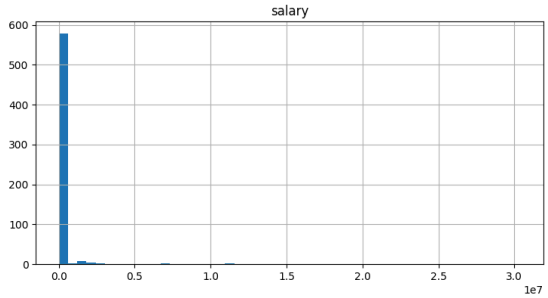
Numerical Features

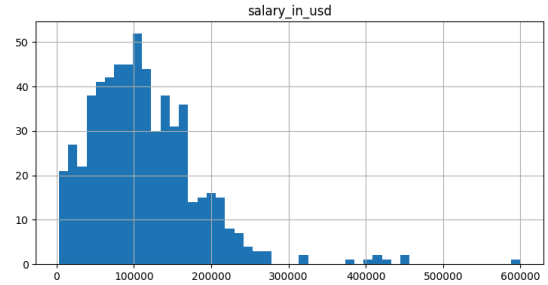
As for the summary and histograms of the numerical attributes:

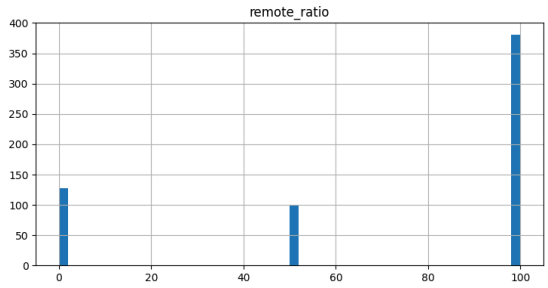
	id	work_year	salary	salary_in_usd	remote_ratio
count	607.000000	607.000000	6.070000e+02	607.000000	607.00000
mean	303.000000	2021.405272	3.240001e+05	112297.869852	70.92257
std	175.370085	0.692133	1.544357e+06	70957.259411	40.70913
min	0.000000	2020.000000	4.000000e+03	2859.000000	0.00000
25%	151.500000	2021.000000	7.000000e+04	62726.000000	50.00000
50%	303.000000	2022.000000	1.150000e+05	101570.000000	100.00000
75%	454.500000	2022.000000	1.650000e+05	150000.000000	100.00000
max	606.000000	2022.000000	3.040000e+07	600000.000000	100.00000











We will drop the `id` and `salary` columns as they are not useful for our analysis. The `salary` is not useful because the data has different currencies so it is not possible to compare the salaries directly. We will use the `salary_in_usd` column instead. The `id` column is not useful because it is just a unique identifier for each row. After dropping the columns, here is the output for `head()`:

	work_year	experience_level	employment_type	job_title	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2020	Mid-Level	Full-Time	Data Scientist	EUR	79833	DE	0	DE	L
1	2020	Senior-Level	Full-Time	Machine Learning Scientist	USD	260000	JP	0	JP	S
2	2020	Senior-Level	Full-Time	Big Data Engineer	GBP	109024	GB	50	GB	M
3	2020	Mid-Level	Full-Time	Product Data Analyst	USD	20000	HN	0	HN	S
4	2020	Senior-Level	Full-Time	Machine Learning Engineer	USD	150000	US	50	US	L

Categorical Features

Next, we will investigate the non-numeric columns. We will use the `value_counts()` method to see the unique values in each column.

```
experience_level
Senior-Level       280
Mid-Level          213
Entry-Level         88
```

```
Executive-Level     26

employment_type
Full-Time    588
Part-Time     10
Contract       5
FL             4

job_title
Data Scientist                          143
Data Engineer                           132
Data Analyst                             97
...
Marketing Data Analyst                    1
Machine Learning Manager                  1
Data Analytics Lead                       1

salary_currency
USD    398
EUR     95
GBP     44
...
AUD      2
CLP      1
CHF      1

employee_residence
US    332
GB     44
IN     30
...
PH      1
NZ      1
CH      1

company_location
US    355
GB     47
CA     30
...
HU      1
HN      1
IE      1

company_size
M    326
L    198
```

```
S     83
```

Based on the results we see from the above, we will drop the column `salary_currency` as it is not useful for our analysis because it is hard to compare salaries in different currencies. We have a column called `salary_in_usd` which is the salary in USD. We will use this column for our analysis. The currency they use may reflect where they reside in or which country they work from. However, we have a column called `employment_residence` which is the employee's primary country of residence. We will use this column to analyze the salaries based on the country of residence.

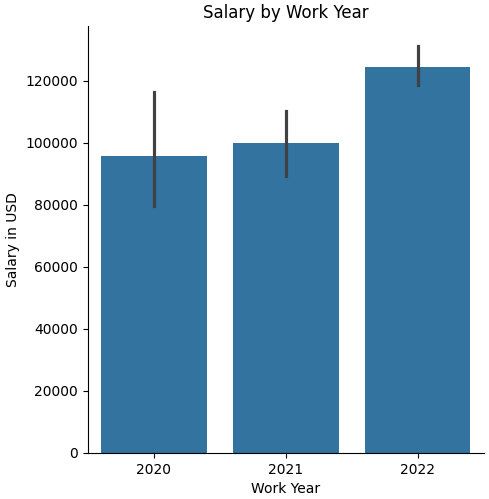
From the above, we can observe that out of 607 entries, there are 588 entries where the employees are working full-time and the remaining 19 entries are working part-time, contract, or freelance. So we can see that the majority of the employees are working full-time. Another thing we can observe is that not all employees are working in the same country as the location of their companies. For example, there are 355 employees working for companies in the US but only 332 employees are residing in the US. This means that there are 23 employees who are working for companies in the US but are residing in other countries. Also, we can see that the majority of the employees reside and work for the companies in the US. The country with the next highest number of employees is Great Britain which has 47 employees.

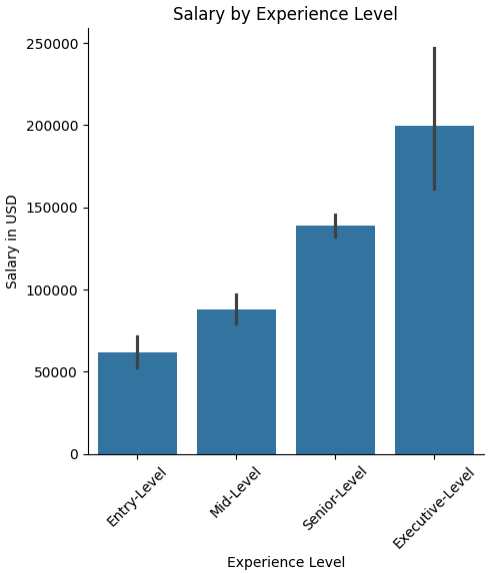
Correlations in the Data

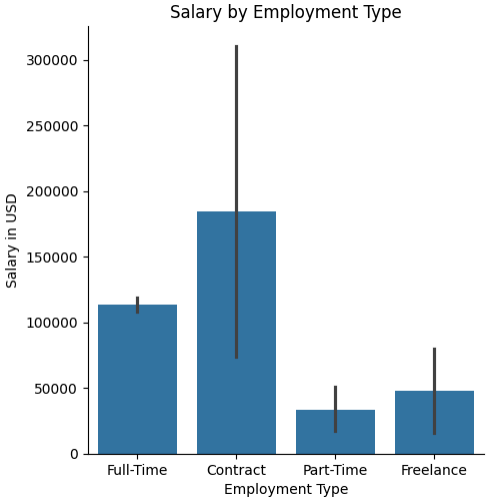
First, we check the correlations between the numerical features:

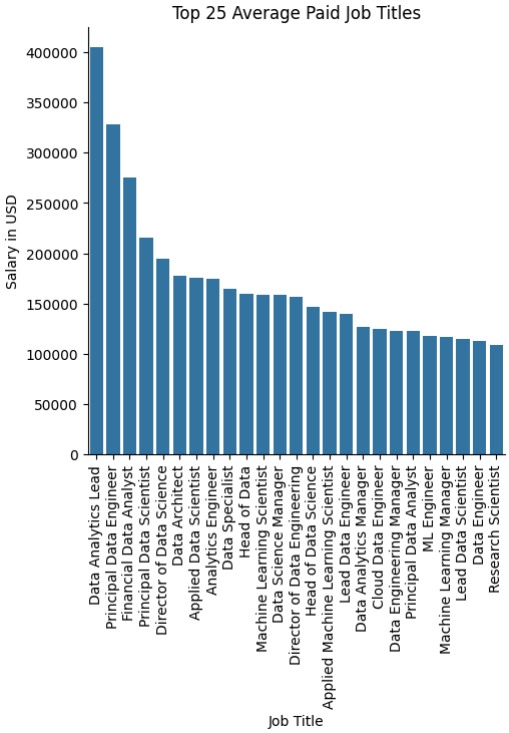
	work_year	salary_in_usd	remote_ratio
work_year	1.000000	0.170493	0.076314
salary_in_usd	0.170493	1.000000	0.132122
remote_ratio	0.076314	0.132122	1.000000

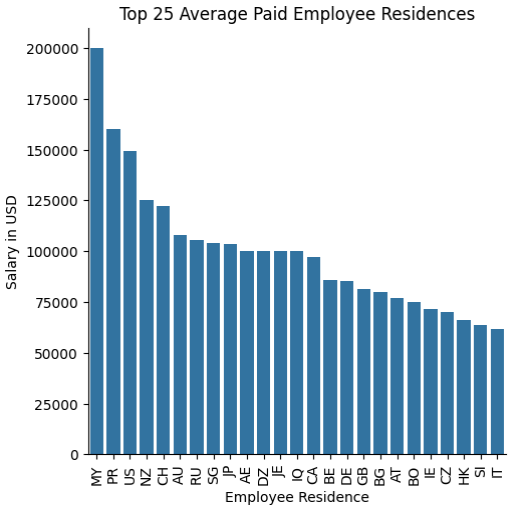
The above correlation matrix does not tell us much of anything, as `work_year` and `remote_ratio` have categorical variables. Instead we plot out `salary_in_usd` versus the other attributes to explore their relationships visually:

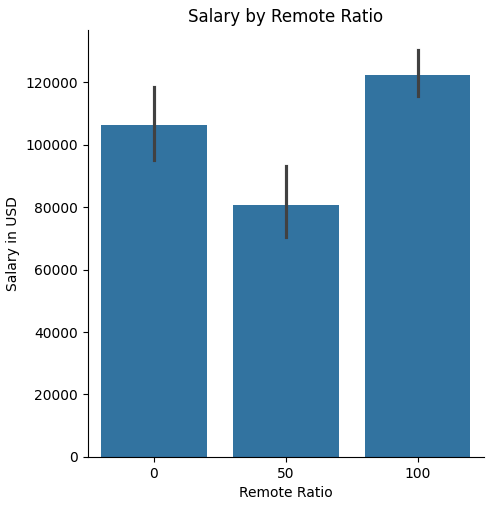

Salary by Work Year
Salary in USD
0
20000
40000
60000
80000
100000
120000
2020
2021
2022
Work Year

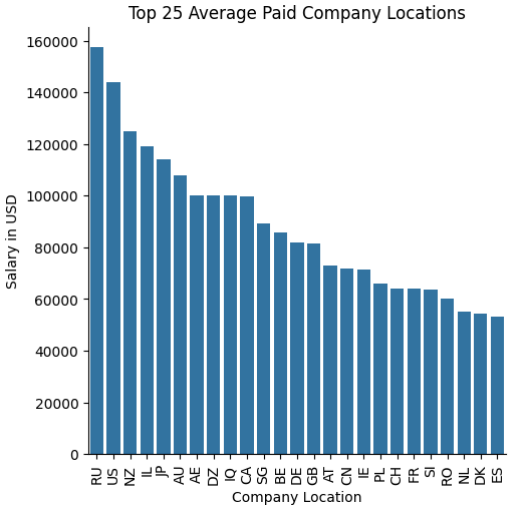

Salary by Experience Level
Salary in USD
0
50000
100000
150000
200000
250000
Entry-Level
Mid-Level
Senior-Level
Executive-Level
Experience Level

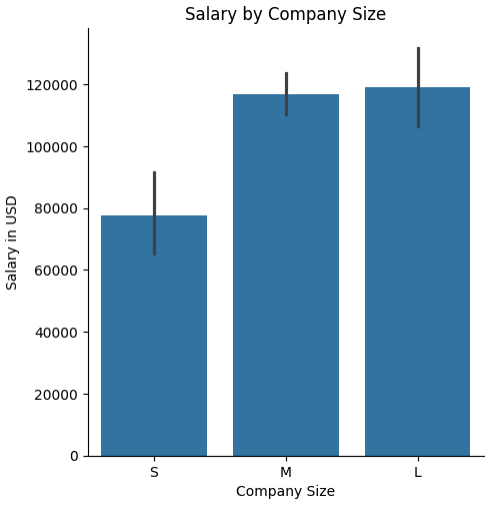

Salary by Employment Type
Salary in USD
0
50000
100000
150000
200000
250000
300000
Full-Time
Contract
Part-Time
Freelance
Employment Type


Top 25 Average Paid Job Titles
Salary in USD
0
50000
100000
150000
200000
250000
300000
350000
400000
Data Analytics Lead
Principal Data Engineer
Financial Data Analyst
Principal Data Scientist
Director of Data Science
Data Architect
Applied Data Scientist
Analytics Engineer
Data Specialist
Head of Data
Machine Learning Scientist
Data Science Manager
Director of Data Engineering
Head of Data Science
Applied Machine Learning Scientist
Lead Data Engineer
Data Analytics Manager
Cloud Data Engineer
Data Engineering Manager
Principal Data Analyst
ML Engineer
Machine Learning Manager
Lead Data Scientist
Data Engineer
Research Scientist
Job Title









Some observations which can be made here are:

- On average, 2022 had the highest salaries for data scientists
- Salary by experience level has a linear increase based on experience
- Contractors have the largest range for their salaries
- Data Analytics Lead job has the highest average salary
- On average, employees who live in Italy have the lowest salary
- For more hybrid work environments, the salary is on average lower
- Companies based in Russia have the highest average salary payouts
- Salaries for medium and large sized companies are roughly the same

With the insights gained from visualizations and observations, human intuition can provide initial cues for salary predictions. However, to enable accurate predictions by a machine learning model, a structured process of data preprocessing, training, and testing is indispensable. This involves transforming categorical variables, splitting the dataset, and training a linear regression model. Through systematic training and evaluation, the model can effectively leverage observed relationships to make informed salary predictions, offering valuable insights for decision-making.

Feature Engineering

Various techniques were employed to transform and engineer features from the raw dataset in order to enhance the predictive performance of our models.

Encoding Categorical Variables

We transformed categorical variables into numeric representations to enable their incorporation into our machine learning models. Specifically, we converted the following categorical variables into numeric form:

- `experience_level`: `('Entry-Level', 'Mid-Level', 'Senior-Level', 'Executive-Level')` → (1, 2, 3, 4)
- `employment_type`: `('Full-Time', 'Part-Time', 'Contract', 'Freelance')` → (1, 2, 3, 4)
- `remote_ratio`: (0, 50, 100) → (1, 2, 3)
- `company_size`: `('S', 'M', 'L')` → (1, 2, 3)

Handling Outliers

We addressed outliers in the `job_title` feature by categorizing infrequent job titles into an `'Other'` category. This transformation mitigated the impact of rare instances on model performance while retaining valuable information from common job titles.

Creating Additional Features

We introduced new features to capture additional information and relationships within the dataset:

- `same_location`: A binary indicator denoting whether the employee's residence matches the company's location
- `mean_country_salary`, `min_country_salary`, `max_country_salary`: Calculated the mean, minimum and maximum salaries within each defined cluster based on the attributes `work_year`, `experience_level`, `employment_type`, `job_title`, `remote_ratio`, `company_size`, and `same_location`

These engineered features aim to enhance the predictive power of our models by capturing relevant information and patterns inherent in the dataset.

Model Selection and Evaluation

We created models before and after feature engineering to determine the most suitable regression models for predicting job salaries.

Before Feature Engineering

Data Preparation and Preprocessing

The dataset's numeric features were scaled using standard scaling, while categorical features were one-hot encoded.

Model Training and Evaluation

We trained the regression models `LinearRegression`, `RandomForestRegressor`, `SGDRegressor`, and `Ridge`.

The models were evaluated by using the mean squared error (RMSE) and mean absolute percentage error (MAPE). In addition, we employed cross-validation with 10 folds to assess model performance.

After Feature Engineering

Data Preparation and Preprocessing

The dataset's numeric features were scaled using `MinMaxScaler`, while categorical features were one-hot encoded.

Model Training and Evaluation

We trained the regression models `RandomForestRegressor`, `SGDRegressor`, `Ridge`, and `DecisionTreeRegressor`.

Each model's performance was evaluated using mean squared error (MSE) and R-squared (R2) scores.

Hyperparameter Tuning

As the `RandomForestRegressor` model exhibited the lowest MSE and highest R2 score among all of the models tested, we performed hyperparameter tuning using `GridSearchCV` to search for the optimal combination of hyperparameters for the model. The following hyperparameters were tuned:

- `n_estimators`: The number of trees in the forest
- `max_depth`: The maximum depth of the trees in the forest

The code with the parameter settings are as followed:

```
rf_params = {
```

```
    'n_estimators': [100, 200, 300],
    'max_depth': [10, 20, 30, 40, 50]
}
```

Results

Metric Evaluations

The following are each of the model's evaluation metrics with BFE being "Before Feature Engineering" and AFE being "After Feature Engineering":

Model	RMSE	MAPE	MSE	R2 Score
BFE: LinearRegression	39274.373	32.528%		
BFE: RandomForestRegressor	54457.504			
BFE: SGDRegressor	53091.896			
BFE: Ridge	53951.563			
AFE: RandomForestRegressor			0.1249169	0.8223989
AFE: SGDRegressor			0.4231541	0.3983790
AFE: Ridge			0.1972537	0.7195537
AFE: DecisionTreeRegressor			0.1435096	0.7959646
AFE: RandomForestRegressor (with GridSearchCV)			0.1227654	0.8254578

RandomForestRegressor (with GridSearchCV) had the best evaluation metrics making it the best model to be used for predictions.

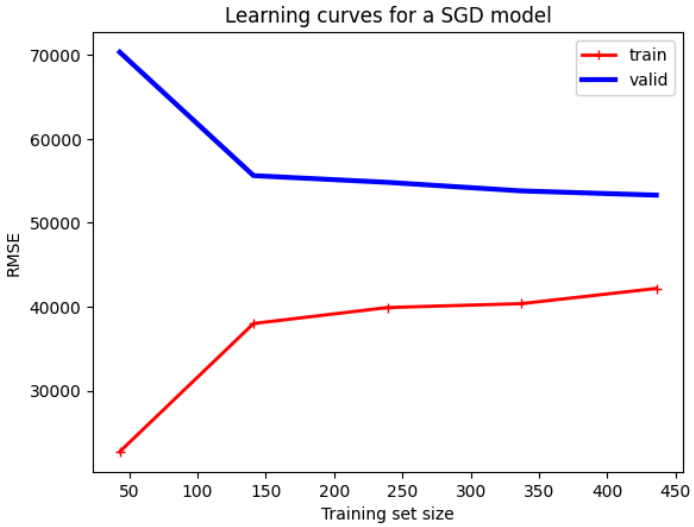
Learning Curves

Before feature engineering, the model which showed the best learning through its training set is the SGDRegressor.

```
train_sizes, train_scores, test_scores = learning_curve(estimator=sgd_reg, X=ds,
y=ds_labels, cv=10, scoring='neg_root_mean_squared_error')

train_scores_mean = -np.mean(train_scores, axis=1)
test_scores_mean = -np.mean(test_scores, axis=1)
```

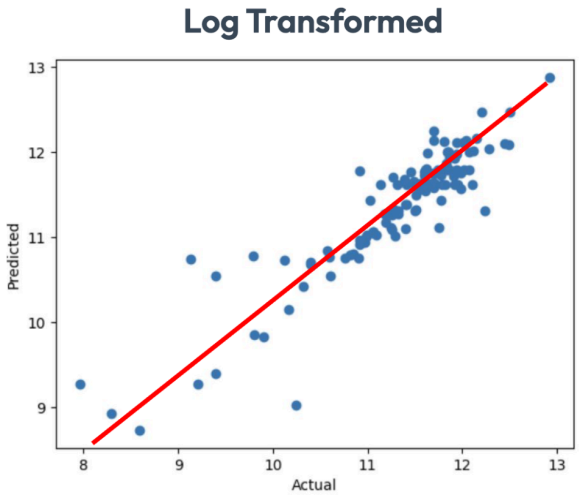
```
plt.plot(train_sizes, train_scores_mean, "r-+", linewidth=2, label="train")
plt.plot(train_sizes, test_scores_mean, "b-", linewidth=3, label="valid")
```

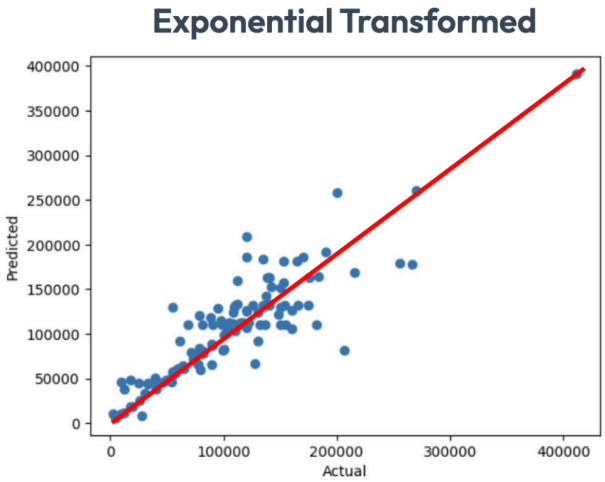


Out of all of the learning curves plotted for the models before feature engineering, `SGDRegressor` is shown to be the closest one to convergence among the training and test RMSE scores. We believe that if we had more data, the lines would be even closer to one another.

Plotting Actual vs Predicted Salary

Here are the scatter plots for the model `RandomForestRegressor` (with `GridSearchCV`):





With the linear pattern across all of our data points, it shows that this model's predictor works.

Conclusion

Throughout this project, we embarked on a comprehensive analysis of data science job salaries in pursuit of greater transparency and equity in data science job markets. Our study delved into the intricate dynamics of salary determinations, guided by our research questions aimed at uncovering notable trends or patterns and whether or not we can create a predictor.

Answers to Research Questions

How do different factors influence job salaries?

Through rigorous analysis, we identified a multitude of factors influencing data science job salaries, including experience level, employment type, geographic location, and the presence of remote work arrangements. Our analysis revealed that one of the most significant findings pertains to the reliance of predictive models on clusters within the dataset. By considering factors such as region, experience level, and employment type, the models can determine mean, minimum, and maximum salary ranges for specific clusters, leading to more accurate predictions. Notably, when examining feature importance, we found that mean, minimum, and maximum salaries exhibited the highest values, underscoring their critical role in the predictive process. Understanding these patterns within the data enables us to better anticipate expected salaries for data science positions.

Are there any notable trends or patterns in salary structures?

Our investigation uncovered several notable trends in salary structures within the data science domain. Analysis of correlations within the dataset revealed variations in compensation based on factors such as company size, job title, and employee residence. For example, we observed that salaries tend to be higher in larger companies and for certain job titles. Additionally, our analysis indicated that remote positions, on average, offer better compensation compared to non-remote roles. These findings underscore the diverse landscape of the job market and highlight the importance of considering various factors in salary negotiations.

If we were to become an entry-level data engineer in 2024, how much would our salary change working in Canada compared to the US? Does remote work play a factor?

To answer this question we inputted the data points with the respective values within the `LinearRegression` model to formulate a prediction:

```
X_new = pd.DataFrame([{
    'work_year': 2024,
    'experience_level': 'Entry-Level',
    'employment_type': 'Full-Time',
    'job_title': 'Data Engineer',
```

```
    'employee_residence': 'US',
    'remote_ratio': 0,
    'company_location': 'US',
    'company_size': 'M'
}])

predictions = lin_reg.predict(X_new)
```

We made further predictions with different company sizes and remote ratios to get the averages in Canada and the United States. Upon doing so, here are the results from our predictor in USD:

	Remote	Non-Remote
Canada	$93,333	$67,000
United States	$161,366	$98,666

Additional Remarks

Our analysis has revealed clear trends in salary structures based on the factors we trained our models on. However, it is evident that the accuracy of our predictor model could be further improved with additional data and a broader geographic focus. By expanding our dataset beyond the US and including more diverse data points, we can enhance the robustness and generalizability of our model, enabling more accurate salary predictions across different regions and contexts.

In conclusion, our study provides valuable insights into the multifaceted nature of salary determinations within the data science industry. By addressing fundamental research questions and uncovering actionable insights, we aim to equip individuals and organizations with the information needed to navigate the evolving job market landscape effectively.

References

- Kaggle. (2022, June 15). Data Science Job Salaries. Retrieved February 26, 2024, from https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries