Employment Salary Prediction Model

Group #1 - Nandor Gallo

Dr. Ruba Alomari

December 4, 2022

1. Abstract

The goal was to completely understand the contributions towards an individual's salary in the data science industry. In order to correctly predict these wages there was an analysis on historical data scientist salaries. The attributes contributing to these predictions were observed to have some sort of correlation with the job salary. These features include, the year the salary was recorded, the title of the job, the actual job salary in USD, and the experience level for the job. The experiment included an analysis on all features. The analysis found that most jobs fall between $60, 000 - $200, 000 for the data scientist industry but this varies with job level. It was also found that the average salary for an entry position was ~$78, 000 and an executive position averages ~$260,000.

2. Introduction

The goal of this project is to create a model for predicting a data scientist's salary based on a set of selected attributes. This can be used by a company or an individual to determine what their wage should be, given their experience, year, and position. As a student, in the field of computer science, it is important to understand each factor contributing to an overall salary. Equally, it is important as an employer to give proper compensation to potential employees. Given the current project's work employers and students can correctly identify starting wages. Entering the workforce an individual will know what their starting wage should roughly look like given their attributes. It is important to know how much your knowledge is worth, you don't want big companies underpaying your worth. Likewise, it is important for companies to know how much an employee is worth based on the attributes. Since our model contains the work year we can determine how the job market increases salary in future years.

Existing solutions focus on all attributes of the data and doesn't give a more tailored response. They also only focus on currency in USD, for Canadians this is a little harder to compare wages. Our model should only focus on the features we're selecting to be predicted. The work will be up to date with current data. The work will also focus on creating models for each individual attribute to get a better understanding on the correlation between the salary and features. Finally, future suggestions will be made to continue research on applications that could affect job salaries. Including, new features that could potentially be recorded to obtain a more accurate result. Potentially merging datasets to create a wider range of data while also looking at potential biases this may impose.

3. Related Work

Salary prediction using Linear Regression [1]

The approach that was used isn't sufficient to accurately predict salary because the only attribute that is being used is experience level. The data should include other attributes for a more accurate prediction. My work will include all features that correlate to salary, giving a more accurate prediction. This work also doesn't include much exploratory data analysis. It only includes number of entries, there needs to be more information on correlation between the attributes and salary. My work includes a statistical analysis on each feature with respect to salaries. Including, average salary for experience level, salaries per job title, and a trend for salaries in the years given.

Salary prediction using machine learning [5]

This work includes a statement that salaries will increase in future years but doesn't give what the actual predicted salaries will be. The work only includes an estimation of future salaries based on the specific year. It doesn't really include what specific job salaries will be so it generalizes job titles. So it's not an accurate representation of what data scientists make. However, my work will make sure to include predicted salaries to specific job titles. Not only job titles, but experience level as well.

How much do data scientists make? [3]

This work is only an exploratory data analysis and it doesn't include a prediction but it includes a good visualisation for the features and how they correlate to the salary. Also, this work doesn't analyze future growth on salaries, just what the current observations are. My work will include a future analysis as well to accurately predict future salaries for upcoming data scientist. This gives the work a purpose rather than just looking at the overall income of data scientists. If you're a current student, you would want to look at future prospects for the industry as well.

4. Experimental Design

Main Objective

To determine a data scientist's salary based on historical data. Predict salaries for each job title. Also, to predict salaries for future years and based on an individual's experience.

Dataset

The dataset contains 11 features and ~1100 entries The feature we are trying to predict is salary in USD but converted to CAD. The data was collected online using an online survey [2]. The information submitted is anonymous.

Method

The features that were established to have a correlation with the target variable were work_year, job_title, and experience_level. In order to obtain a better understanding between the features and salaries. We will produce a linear regression model for each feature and then combine all features for our final model. This gives us a better outlook on what effects data scientist job salaries. If the data isn't sufficient for the model, an ensemble model will be made to reduce error. In order to increase the model's accuracy extreme gradient boosting will be applied

Baseline

- Provides a better insight to features with respect to job salaries because it works with each independent feature and than gives a final model with respect to each.
- Data is up to date
- Convert USD salary to CAD for a better understanding to Canadians.
- Looking at mean salaries per experience level, job title, and year will give us a better indication of our model results
- Provide future insights that weren't established in previous work.

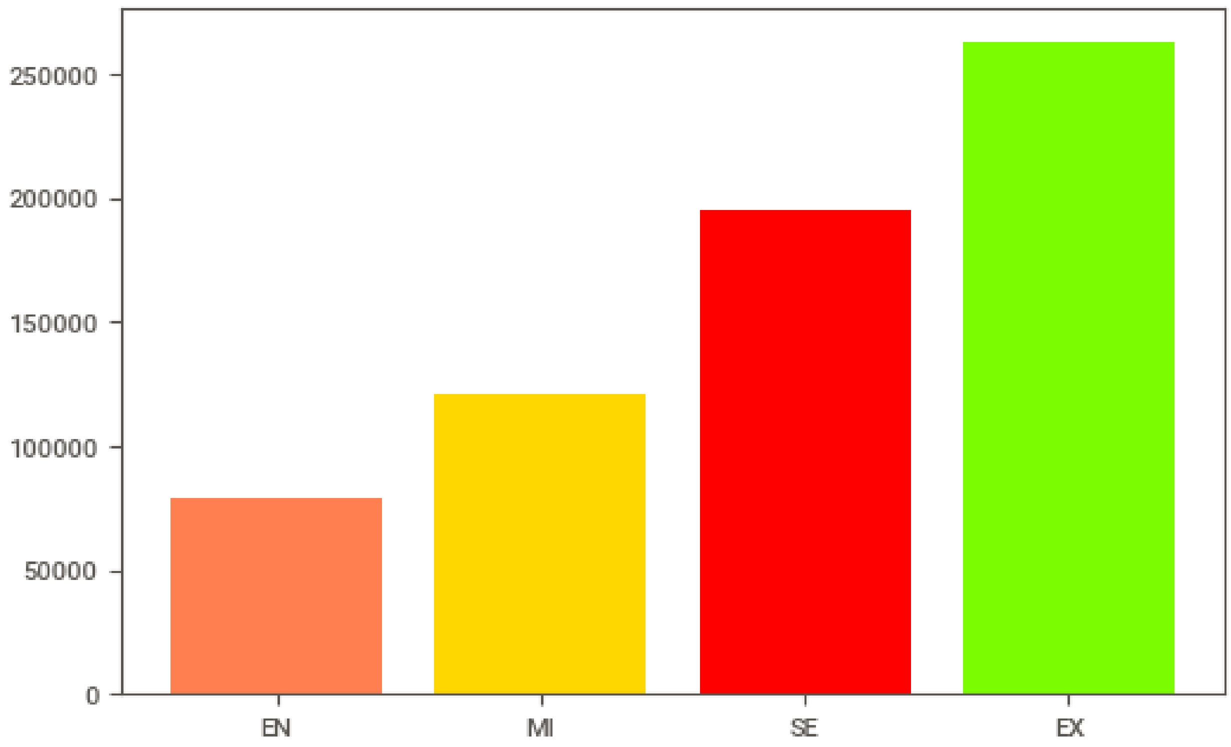
Evaluation Metrics

To evaluate performance, we will take the Root Squared Mean Error of our estimations. This metric is preferred because the unit of RSME will be the same as the predicted unit making it easier to interpret the results of the model.

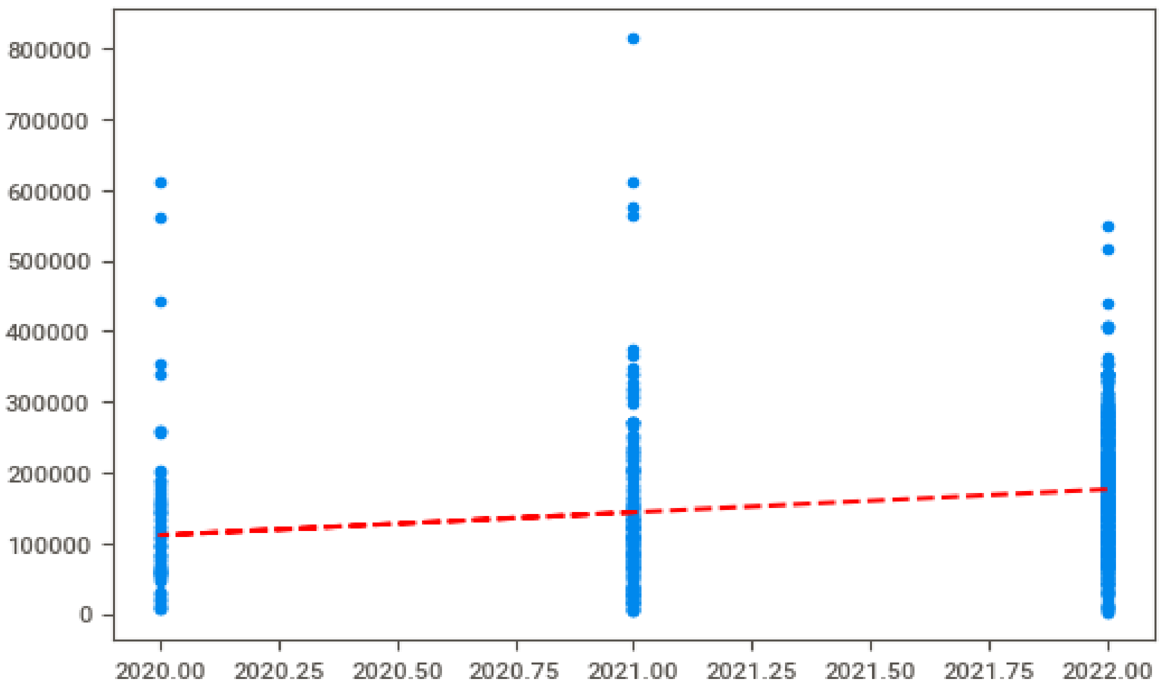
5. Experimental Results

Main findings

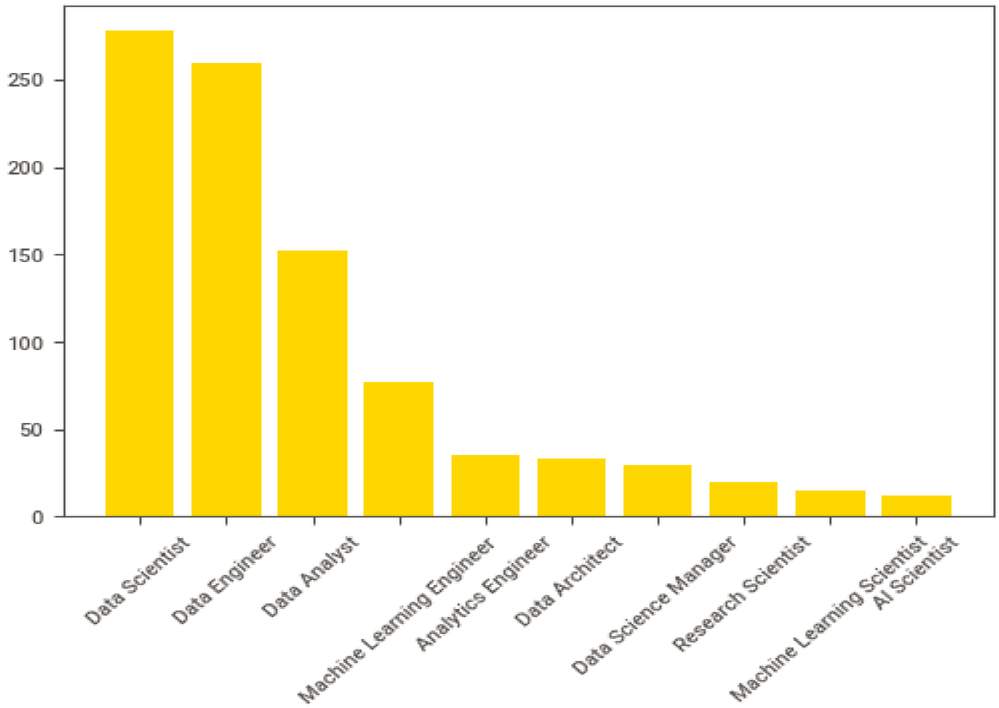
The figure shows a clear distinction in average salaries per experience level. There is no dispute that the more work experience an individual has will lead them to a greater salary. Given this information our model will most likely predict a greater job salary with a greater experience level. This conclusion makes sense because a company will pay you more given your skills, so more experience will lead to a higher salary.



The red line represents the line of best fit for salaries. The x-axis represents years. Here we can see as the years increase so do the salaries. This can give the implication that as time goes on wages increase. This results can be due to economy inflation where prices gradually increase over time, so wages have to increase due to the economy. On the other hand, this can just be a data bias because we could have more higher paying job entries in later years of the data being collected. More work can be conducted to see if this is true or a sampling bias.



The bar chart shows which job positions our data is coming from. We can see that data scientists make up a majority of the data samples. This is prevalent because the data being collected is intended for data scientists, so when surveying the population people most likely just put data scientist instead of their detailed job title.



6. Future Work

For future research we could observe as to why future years show an increase in salary. This could be a false correlation because it could mean more entries of higher paying jobs were entered during those years. Likewise, this can also show an increase in salary due to the economy of the country. So, with the addition of more research we can obtain more knowledge as to why salary increased with these years. Next steps, to continue collecting data on data scientist's salaries and to see if the conclusions drawn are true. The continuation of data collection will hopefully bring out the data's true insights because in statistics the law of large numbers will always produce the expected outcome.

Future work would also require a wider range of features to get a more accurate salary prediction, I believe everything that contributes to a salary isn't included in this dataset. But a good portion is here. A good idea for future work would be to either collect more varying features that contribute to a salary, such as benefits being included, specific company, and years worked at the specific company are some examples of more attributes. Alternatively, we could include another dataset that contains job salaries and different features but that would include some sample biases. So the best option would be just to collect more features while surveying.

7. Conclusion

The key insight this model provides is the correlation between salaries and the selected features. These features are the major contributing factor for determining a job salary. The model is supported by the analysis. We observed that future years and greater experience lead to a higher salary. Using this model will provide a greater insight to the industry with respect to salaries and the given attributes. Also, the model provides a future analysis which is important to future data scientists and employers.

An ethical issue the work can impose on are a breach of privacy on an individual's salary. If someone is found to have a greater salary than you for producing similar work could lead to unnecessary problems at work. Fortunately, the data being collected is anonymous. However, if future work is to be done there would be more features that are more specific to certain individuals such as, the company name someone is working at or the years they worked at that company. Through a deduction process you could figure out an individual's identity.

Another issue with the work is transparency with the data being collected. It is anonymous so what stops one person from entering multiple entries. This can lead to outliers and a greater bias in the data. Since this is intended for students and employers, this can potentially mislead them for either the better or worse.

References

[1] Đức , D. (2022, September 15). *Salary prediction using linear regression*. Kaggle. Retrieved October 26, 2022, from https://www.kaggle.com/code/ducduong18/salary-prediction-using-linear-regression

[2] *Ai-jobs.net salaries*. salaries.ai. Retrieved October 26, 2022, from https://salaries.ai-jobs.net/

[3] Aliphya, G. (2022, June 22). *How much do data scientists make?* Kaggle. Retrieved October 27, 2022, from https://www.kaggle.com/code/aliphya/how-much-do-data-scientists-make/notebook#Do-remote-employees-face-a-pinch-in-their-salary-due-to-their-location(not-working-for-ofc)?

[4] Manral, M. (2022, September 11). *Employ Earnings Data*. Kaggle. Retrieved October 1, 2022, from https://www.kaggle.com/datasets/e1cfbb38c0fe2129a6e744aff1ebd180d4d4c8097a17f9f2860027c0c0793c36

[5] White, L. (2022, August 12). *Linear regression 401d19*. Kaggle. Retrieved October 26, 2022, from https://www.kaggle.com/code/lieslwhite/linear-regression-401d19