Salary Prediction Utilizing Random Forest Regressor Model

Spencer Moore, Kylie Tse, Advait Chalke, and Chiyou Xiong
Department of Computer Science
University of California - Davis, Davis CA, USA 95616
Email: {samoo, kktse, achalke, chixiong} @ucdavis.edu

Abstract — The continuous evolution of the data science field has developed a diverse landscape of job roles and corresponding salaries. Due to this, understanding the determinants of these salaries for each position is crucial for both seekers and employers. The objective of this paper is to explore the predictive capabilities of a model utilizing a comprehensive data science salary dataset in order to forecast potential salaries. Our approach involved cleaning the data of any outliers and isolating features of the data that we determined would best aid in predicting average salaries. The model we determined would perform best with the complexity of the data was the Random Forest Regressor Model. The results of training from the model resulted in an R^2 score of 0.548 which suggests more data manipulation or computational resources are necessary to improve the model but looks promising in its predictive capabilities of salaries in data science.

1 Introduction and Background

In today's competitive job market, being able to predict salaries is very helpful and important to those looking for a job, especially for those that are looking for their first job. From the start, the ability to predict salaries makes people aware of what they should be making, enabling them to negotiate for better salaries, comparable to their peers of similar background in the field. Thus, the ability to predict salaries offers transparency for potential employees, preventing them from being mistreated before even their first day of work. For employers, this transparency also allows them to reconsider their offers to employees, wanting to make theirs better than average to secure workers. Additionally, for those thinking about certain jobs, they can see if they would be satisfied with the average salary before pursuing their career.

The problem we are addressing through our project centers on the ability to develop a rigorous salary prediction model utilizing a dataset. This dataset itself is composed of various attributes of over six hundred individuals ranging from columns labeled by experience level, job title, company location, salary, and so on. Through the analysis of the data using ML techniques and methodologies, we hope to isolate attributes that best predict the salary and then not only relay what attributes individuals need to hone in order to receive a better salary but also inform them of the proper baseline salary they should be receiving from a potential employer. By properly predicting the salary given by the associated attributes, we hope the utility of our project allows individuals to better negotiate for the average salary expected in their field as well as inform them of the various salaries across the field of data science careers.

2 Literature Review

The idea of predicting salaries for workers is not new, and has been covered in research before. However, our project aims to focus specifically on workers in data science, since a more specific model will have a better chance of accuracy. Several academic approaches have been taken towards predicting salaries, which gives our group a foundation to begin upon. For instance, many regression models have been used[1][2]. Although another article reached the conclusion that another model type, a "decision tree model", is better suited for the task than regression[3]. Using a regression model has the benefit of relying on research already done. While reviewing the academic

background of the problem, these salary prediction models tend to focus on a wider range of audiences and careers, such as census data across thousands of jobs[3]. By focusing on careers and salaries in data science specifically, while still using regression, our team has a plethora of research to depend upon for the regression model while still developing a new and useful model. While we acknowledge that salary prediction is a difficult task, by focusing on a specific field we hope to have greater accuracy.

3 Dataset Description and Exploratory Data Analysis

There are a total of 12 columns and 607 entries (data points) ranging from 0 to 606. The data type is either integer or objects. The columns indicate a persons work year, experience level, employment type, job title, salary, salary currency, salary in usd, employee residence, remote ratio, company location and company size.

	Unnamed: 0	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

Fig. 1 Statistics of the dataset.

Some salaries are in different currencies and so the best way to go around that will be to look at the data corresponding to salary in usd for a standardized data set. In Figure 1, we can see that the mean salary is 112297.87 with a standard deviation of 70957.26. Moreover, the count for the columns is 607 implying that there is no missing data.

In Figure 2, we can see that there are quite a few outliers in the salaries (in usd), around 300000 and higher. When training our data, we will want to remove these. Since the mean is greater than the median, the distribution of the salaries is right-skewed. This means that most salaries are below the mean of 112297.87.

In Figure 3, we see that those that are considered

Fig. 2 Box plot of average salaries in usd.

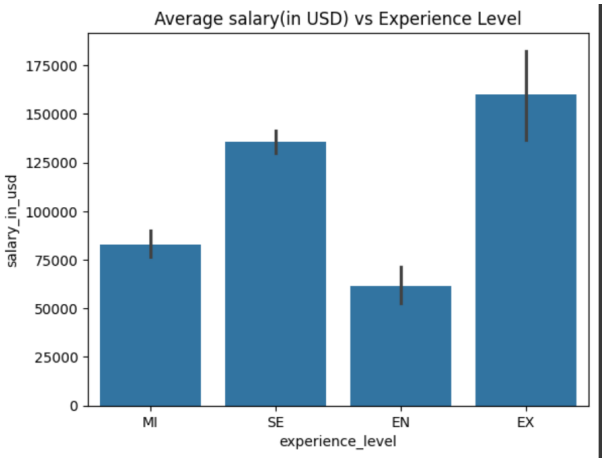

Fig. 3 Bar plot of average salary vs experience level.

to have executive level experience had the highest average salaries, at around 200000, and those that are considered to have entry level experience had the lowest average salaries, at around 60000. Those with senior level experience had the second highest average salaries, followed by those with mid level experience. We expected as such, since more experience should generally mean higher pay.

In Figure 4, we notice that the average salary has been increasing in recent years. In 2020, the average salary was only around 80000, but in 2022, the average salary jumped up to around 120000.

In Figure 5, we see that medium sized companies had the highest average salaries, at around 118000. Large sized companies had average salaries of around 115000. Small sized companies had the biggest difference in average salaries, at a low of about 70000.

In Figure 6, we see that employees that reside in the United States have the highest average salaries

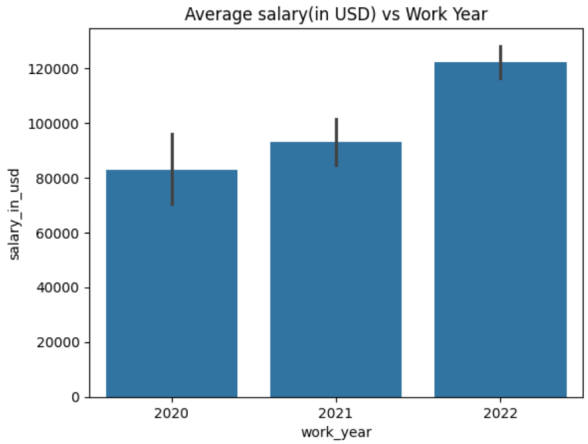

Fig. 4 Bar plot of average salary vs work year.

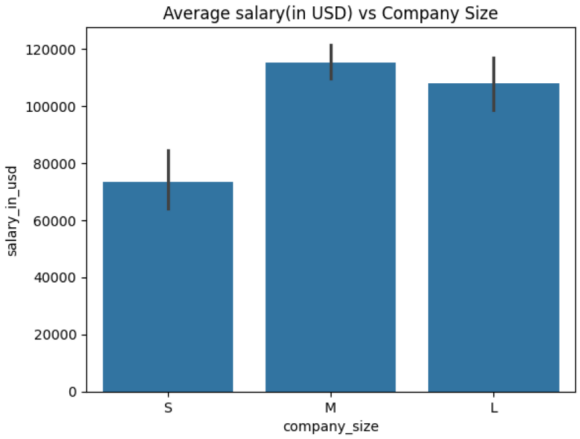

Fig. 5 Bar plot of average salary vs company size.

of around 140000, followed by those that reside in Japan, with average salaries of around 100000. The two employee residences with the lowest average salaries are Portugal at around 48000 and India at around 40000.

4 Proposed Methodology

Each sample was complete without missing values, which meant we did not have to remove samples for being incomplete. We picked out specific features from the dataset to use in our model, which included experience level, remote ratio, company size, job title, company location, and employment type. We removed only the upper outliers from the dataset. While experimenting with the model initially, we noticed that it really struggled to predict excessively large salaries. And since those excessively large salaries were outliers anyways, it felt best to just remove them.

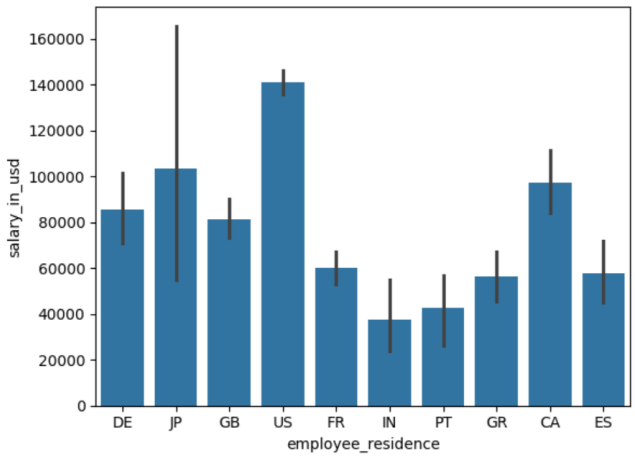

Fig. 6 Bar plot of average salary vs employee residence.

We used libraries like sklearn and numpy to create our model. To set up our data, we used the previously mentioned features as our X variable, and the average salaries in US dollars as our y variable. We used one hot encoding for the categorical features, which included experience level, company size, job title, company location, and employment type. The dataset did have primarily categorical features, which poses a unique challenge for machine learning models. But by using one hot encoding, this converted the features to have numerical values of 0 or 1. Then we standardized both the features and the target variable using the Z-Score method, even though most of the inputs were already between 0 and 1. Finally we split the data into training and test sets, using a ratio of 75:25.

Next, we set up our model using a Random Forest Regression Model since the relationships between our data are more complex, and the model should perform well. We performed hyperparameter tuning using grid search on various parameters such as: the number of estimator trees in the forest (130, 140, 150), criterion (squared error, absolute error), maximum tree depth (11, 13, 15), minimum number of samples to split an internal node (4, 6, 8), and minimum number of samples to be at a leaf node (4, 6). We actually did the grid search multiple times, changing the possible parameters each time until we reached the optimal set of parameters. By doing grid search iteratively on smaller sets of parameters,

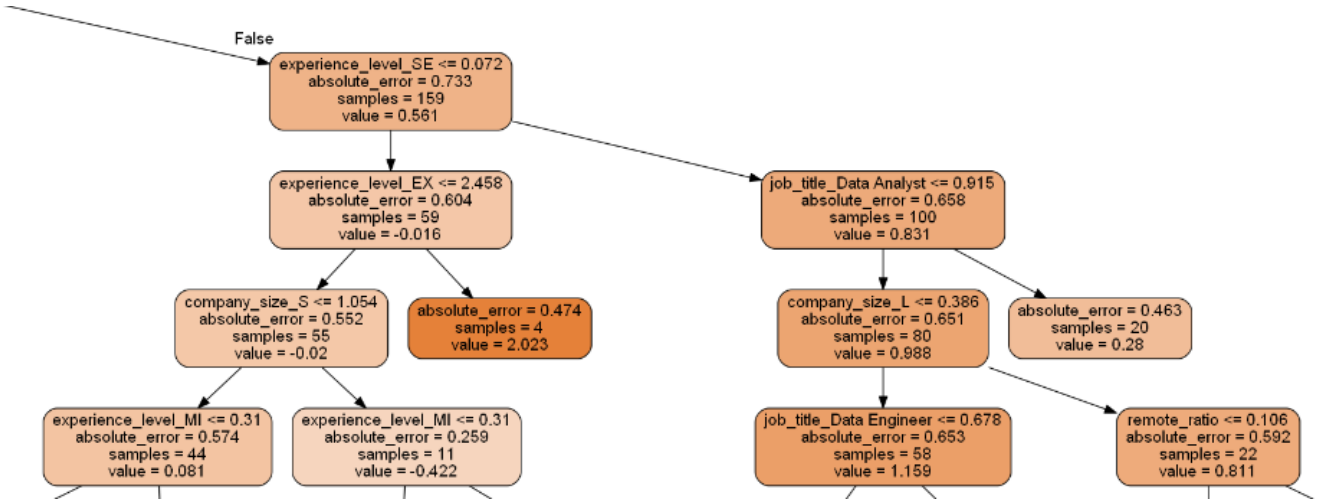

Fig. 7 Part of the first estimator of 130 in the final Random Forest model.

we could find the best parameters much faster than searching over a massive grid. Once finalized, we used the best hyperparameters to train our model and make predictions on the test set. To evaluate the performance of the model, we computed the mean squared error and the R-squared scores.

5 Experimental Results and Evaluation

After running our model and using grid search, we found the best hyperparameters to be a criterion of absolute error, a maximum tree depth of 11, a minimum number of samples to split an internal node of 4, a minimum number of samples to be at a leaf node of 4, and a number of estimators of 130. We used these hyperparameters to train our final model, which has an MSE score of 0.615 and and R^2 score of 0.548. An R^2 score of 0.548 means that approximately 54.8% of the variance in the target variable can be explained by our model. A section of one estimator is shown in Figure 7, as an example of how the trees ended up looking. The estimator visual was generated with graphviz in Python.

To evaluate how our model performed, we looked at 100 outputs (shown in the submitted python notebook), and some of the predicted values were decently close to their respective target values. Thus, our model is effective to use as a start, although it can still absolutely be improved on. While looking at all non-outlier samples, the model does not perform great overall, as shown in Figure 8. The scatter plot was created using matplotlib in Python. The samples have been sorted based on the target salary in US dollars, and the index is just the sample number in this list.

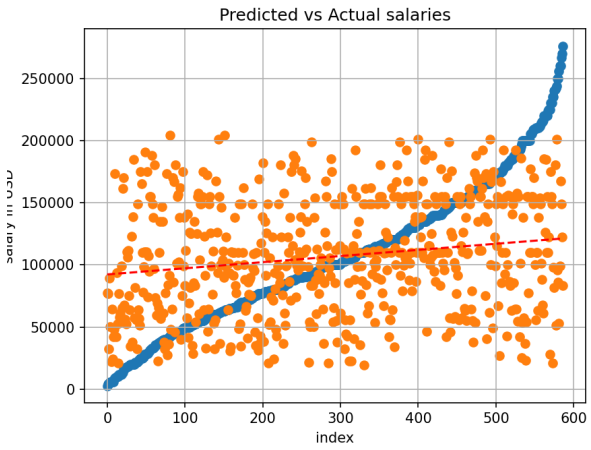

Fig. 8 A plot of the model's predictions, sorted by the target salary. Blue is target, orange is the model prediction, and the red line is the prediction trend line.

While the predictions do technically trend upwards, the model is not as accurate as we would like. This is likely due to a combination of faults, but the problem is also not an easy one to solve. Each company pays differently, and it is very difficult to predict an accurate salary just based on location or job title. The model has learned several trends in the data though, such as realizing that a higher experience level correlates with increased salary. As such, the model still has value as a tool for data science workers looking to

improve their salary. The demonstration video of the model showed how a person could use the model; by inputting their career characteristics and then modifying certain parameters to see how the expected salary changes. Ultimately the model did not perform as well as we had hoped, but it still shows promising signs of being a useful tool if given more development and time.

6 Discussion

The R^2 score of our current implementation of the Random Forest Regressor Model suggests that the predictive accuracy is at a moderate level, with still significant variability unaccounted for. While we found that some of the more influential predictors were experience level and country, there is still more work that could be done to increase the model's predictive accuracy.

In regards to existing literature, our findings are consistent with research that highlights the utility of the Random Forest Regressor model to predict salaries but differ in robustness due to our lower R^2 score and general inaccuracy comparatively. While other literature with high accuracy includes features like specific skills alongside job titles, the dataset our model was trained on has more categorical values most likely resulting in an error on our part in failing to properly isolate parameters that best serve as predictors as well as potentially unintentional constraints from preprocessing steps such as one-hot encoding which may not fully capture the nuances of the data.

In the future, we can enhance our model by refining data pre-processing methods and investigating other methods that may better suit the model. Furthermore, we could better hone which parameters from the dataset best contribute to predicting salaries instead of all the variables currently present in the model.

In relation to the utility of our topic, the ability to predict salaries with reasonable accuracy has significant practical applications for employers and job seekers. This applies to not only setting realistic expectations but also allowing room for salary negotiation. This transparency of salaries helps individuals ensure they are compensated fairly while for the employers this model gives insight into market salaries which can aid in designing positions to attack and retain talent in the data science field.

7 Conclusion

We found that a Random Forest Regressor was useful in predicting employee salary based on given attributes of the dataset. We used One-Hot encoding and a Z-score scaler, as well as removing outliers, to make the data usable and feed it to the Random Forest Regressor model. With hyperparameter tuning through a limited grid search we found the best hyperparameters for our dataset. After training out model, we could predict employee salaries with an R^2 score of 0.548. By better scaling the data into usable values and by having more computational resources, we believe that we can better improve the accuracy of the model. Additionally, while we performed analysis to select what seemed to be the most useful features of the dataset, it is likely that some of the chosen input features are too complicated to be effectively used in a simple model. For instance, while job title is a strong indicator of salary, it is also very nuanced and still dependent on other factors.

8 Additional Information

8.1 Github Link

https://github.com/Orpheus-1/ECS-171-Group-Project

8.2 Project Roadmap

- Choose and understand the dataset - done by week 1
- Describe the problem scientifically - done by mid quarter
- Study related work - done by mid quarter
- Perform exploratory data analysis - mostly done by mid quarter
- Develop the model - done by end of quarter
- Train and evaluate the model - done by end of quarter
- Make the demo - done by end of quarter

- Finish the paper - done by end of quarter

8.3 Assignment of Tasks

Items not specifically mentioned were worked on by everyone.

- Spencer: model training, code demo, results and evaluation, literary review
- Kylie: exploratory data analysis, methodology, results and evaluation
- Advait: exploratory data analysis, model training, conclusion
- Chiyou: abstract, introduction, discussion

References

1. Das, S., Barik, R., and Mukherjee, A., “Salary Prediction Using Regression Techniques” (January 28, 2020). Proceedings of Industry Interactive Innovations in Science, Engineering Technology (I3SET2K19), Available at SSRN: https://ssrn.com/abstract=3526707 or http://dx.doi.org/10.2139/ssrn.3526707
2. Matbouli Y.T., and Alghamdi S.M., “Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations”. Information. 2022; 13(10):495. https://doi.org/10.3390/info13100495
3. Asaduzzaman, A., Uddin, M. R., Woldeyes, Y., and Sibai, F. N., "A Novel Salary Prediction System Using Machine Learning Techniques," 2024 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT NCON), Chiang-mai, Thailand, 2024, pp. 38-43, doi: 10.1109/ECTIDAMTNCON60518.2024.10480058.

1
2
3