IT job Market Analysis and Salary Prediction

# Introduction

This project is primarily focused on IT job market analysis, which is a rapidly growing industry. This project aims to analyse the market trends, its volatility and also discover things that cannot be found at a single glance at the data, the data for the project was collected from various job posting platforms like LinkedIn, Glassdoor etc. Additionally, the project also predicts the salary of various jobs in different fields in IT based on the skills one possesses as well as the experience and qualification and the area where the person is looking for job along with some other features.

# Related Work

A related work that can be compared to our project would be Data Science Job Market Analysis done by Ai Zhong (link: <https://medium.com/@aizhong.ds/data-science-job-market-analysis-9fecc7518230>). Her project goal and ours strongly resemble each other’s since both wanted to analyse IT sector and wanted a clear insight towards the job market. While she focused only on her field which is Data science, we had a much broader point of view which was to consider the entire IT sector. Also, her work was based upon data which was collected in 2019 – 2021 while our data has more latest data which is from 2021-2023.

While her project was based upon a categorical target variable ours is on a numeric target variable which is in a way different, what she is trying to predict is in demand skills while we are trying to predict average salary.

Even so, while comparing the benchmarks their model has an F1 score of 0.71 and our R2 score is 0.88 while these values cannot be compared with each other there are some insights that could be drawn from this. The F1 score of 0.71 would be a decent score which would mean how well one can classify while a R2 score of 0.88 can be considered excellent.

Other works that can be compared to ours would be works done by the other users in Kaggle, so far no one was able to completely build a salary prediction model, at best they could achieve was an incomplete analysis, though there are some who has done an entirely different project which cannot be compared to ours.

# Methods

### Brief Overview of the data

The dataset has data starting from 2021 to 2023 with 23 columns and almost 10 million rows, which is comprised of mostly categorical features. The most important features experience range, salary range, location, skills, qualification, work type, job title, company and its size, preference of gender etc. Most of the features are string while the others are in a range format. The data set was downloaded from Kaggle (<https://www.kaggle.com/datasets/ravindrasinghrana/job-description-dataset>).

## EDA

### Data cleaning

* The initial data set with 10 million rows had all the jobs posted on various platforms which were unnecessary for our project, we filtered out the IT job roles manually by going through all the job titles and the result was a 3 million row dataset.
* Then all the features were renamed so that it may be easy to use later on.

### Feature Engineering

* The feature salary range was split in to two minimum and maximum salary and then calculated the average salary based on the minimum and maximum.
* The feature skills were in text format which was then reconstructed to have only 4 skills based on the key words in the text.
* The feature experience had a range value too (e.g.: 0 to 3 years), this was remedied in the same way as salary range, by splitting them apart and creating a minimum and maximum experience and then creating a new column which is average experience.
* For easy analysis we grouped the country column to represent its respective continent and created a new column called continent.
* Also created a new feature called Industry based on the job title we created this feature which would represent the job in a broader way (e.g. Back-end and front-end developer both came under the industry of Software development).
* Created some categorical features using the existing features such as salary category and experience group.

### Statistical Hypothesis

Did several tests to find out relation between the features and its values. Here are some of the tests that were conducted:

* Mann-Whitney U Test: we used this test because the data was a bit right skewed as well as not normal. The test was conducted on the salary distribution of two different industries.
* Chi-square test: We assessed the relation between the salary and company and got a p- value of 0.8 which is higher than 0.05 which would suggest that there is no significant association between salary and company.

### Correlation Heatmap

The correlation matrix of the numeric features was found and a visual representation in the form of a heatmap was also drawn as a result.

### Visualisation

The data analysis with the aid of visuals such as graphs were done in python as well as using Tableau, which is extensively used for data visualisation. 2 to 3 dashboards were depicted which would say the whole story of the data.

### Scaling

* Logarithmic scaling was used on minimum and maximum salary as well as on experience, this makes the low values more expanded at the same time make the high values to shrink.
* Standard Scalar was used on the target variable which is average salary which was in the range of 65,000 to more than 100,000 at times. Scaling was done in order to help the model train well as well as easy to analyse the results.

### Encoding

* Due to the high number of unique skills, the values in this feature were mapped to various categories such as skills like python, java, SQL, JavaScript were all incorporated to a single category as programming.
* Target Encoding was also employed in the case of job title in order to reduce the dimensionality of the data after encoding. This particular type of encoding lets the feature values to stay in the same column while it assigns a value to each of the value in the feature based on the target variable.
* Label encoding was used on features like experience group and salary category.
* One-hot encoding was used on features like continent, work type and preference.

### Feature Selection

Two methods were used for this Wrapper method and Embedded method.

* Embedded method creates a random forest regressor and then calculates the feature importance.
* Wrapper method runs a linear regression model and eliminates the features that are less important thus it is also called recursive feature elimination.

### Pycaret

Pycaret was employed to compare regression models, both the features that were selected from the wrapper method and embedded method were used separately. The best model from both of them were selected to continue on to fine tuning.

### Machine Learning model

Gradient Booster regressor was employed as it was the best model that yielded the highest R2 score. The features selected from the embedded feature selection were used for this model.

### Hyper Parameter Tuning

The hyper parameters that can be changed and hence tuned manually were implemented in order to improve the accuracy of the model.

### Deployment

The model was deployed in gradio, gradio is a tool which let sone build a user interface for the machine learning model. After running the code, the link for the interface will be visible immediately and as long as the platform where the model is running is not shut down the interface can be accessed and it will predict the values using the model in the back end.

# Results

### Data cleaning and Feature Engineering

After cleaning there were only about 3 million jobs, and also removed any null values present. Feature engineering made the data more interpretable, specially creating average salary and average experience will help greatly in the model building as well as analysis.

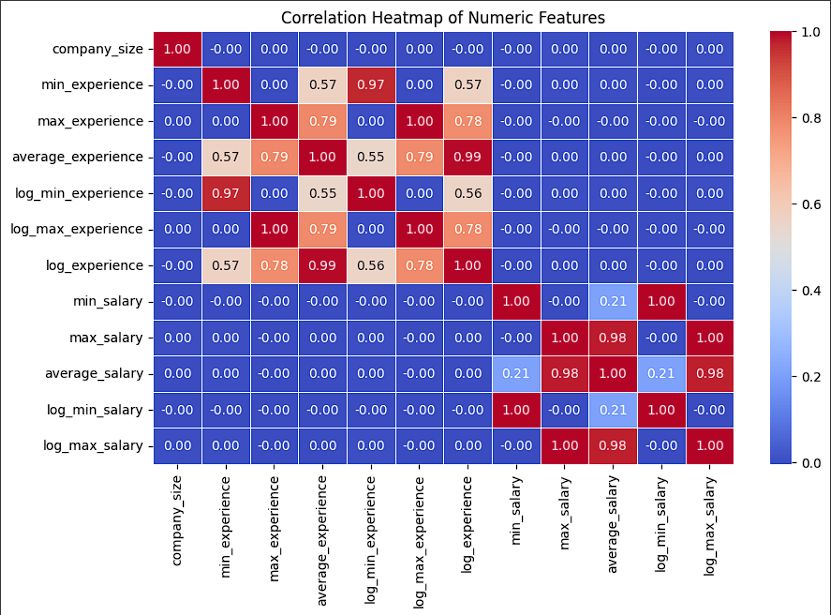
This is what the features look like afterwards:



### Statistical Hypothesis

* Mann-Whitney U test: The result as a p-value of 0.02 which is less than 0.05 so we reject the null hypothesis which state that there is no difference in the salary distribution across the industries UI/UX design and software development.

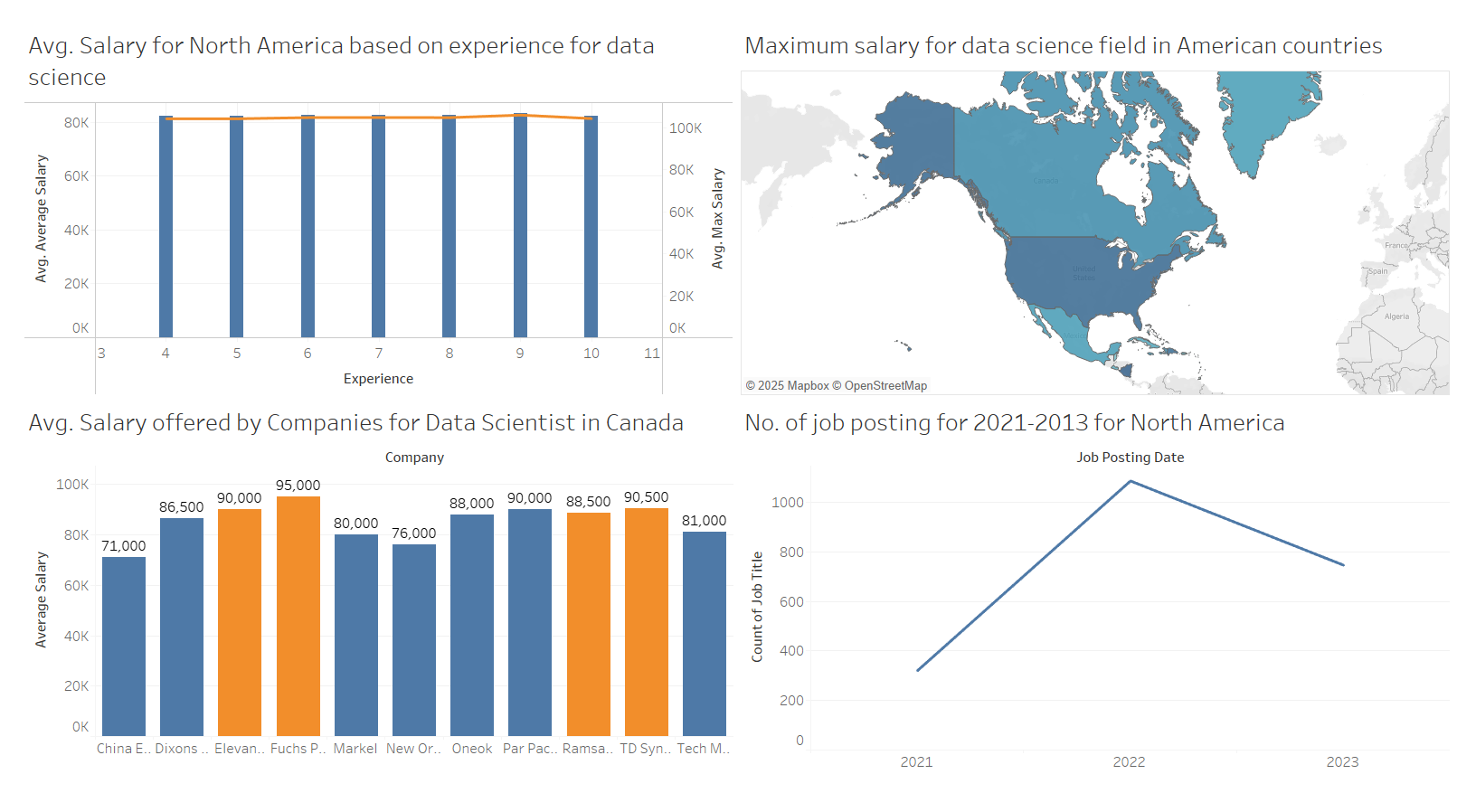
### Correlation Heatmap

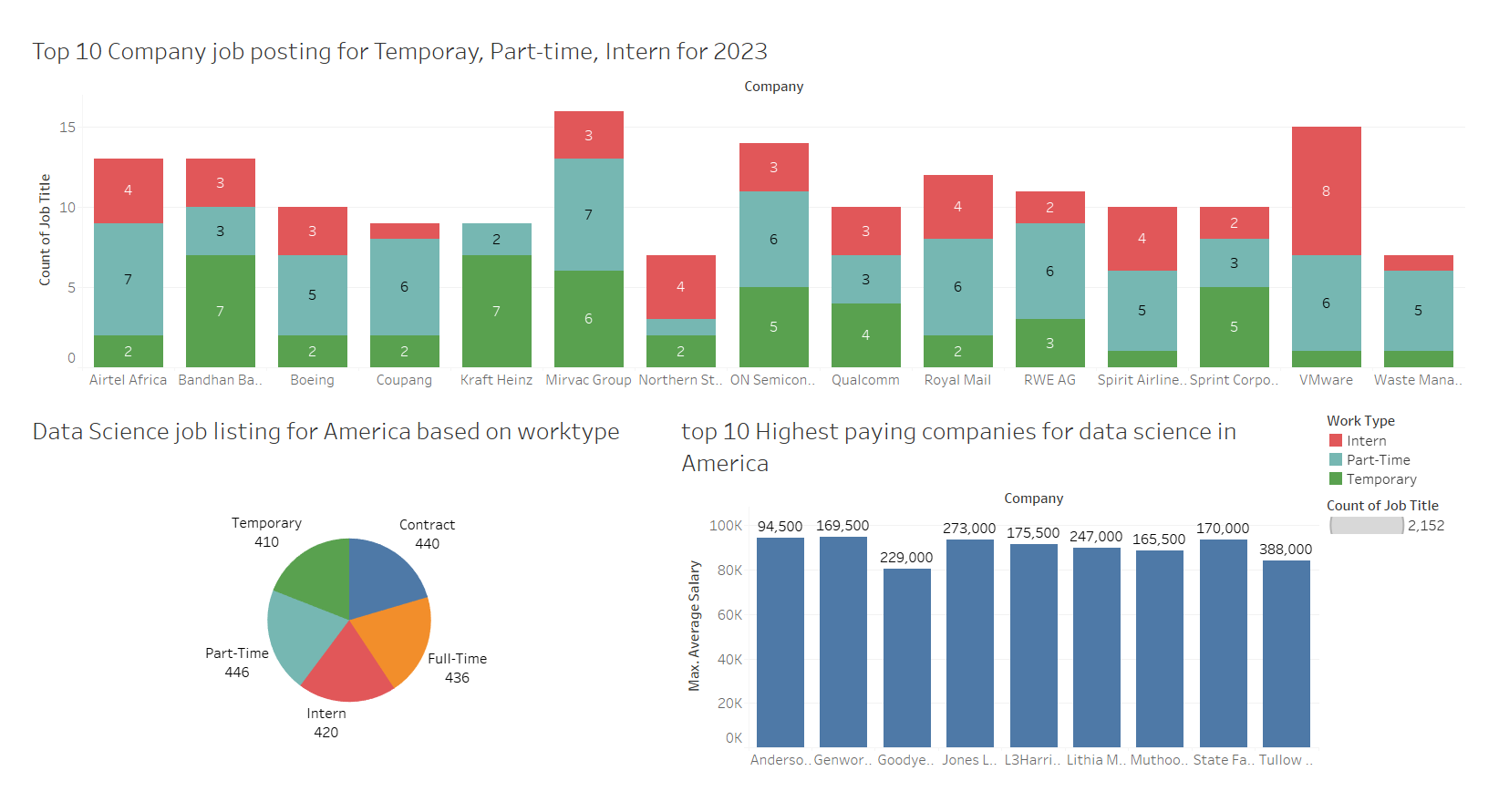


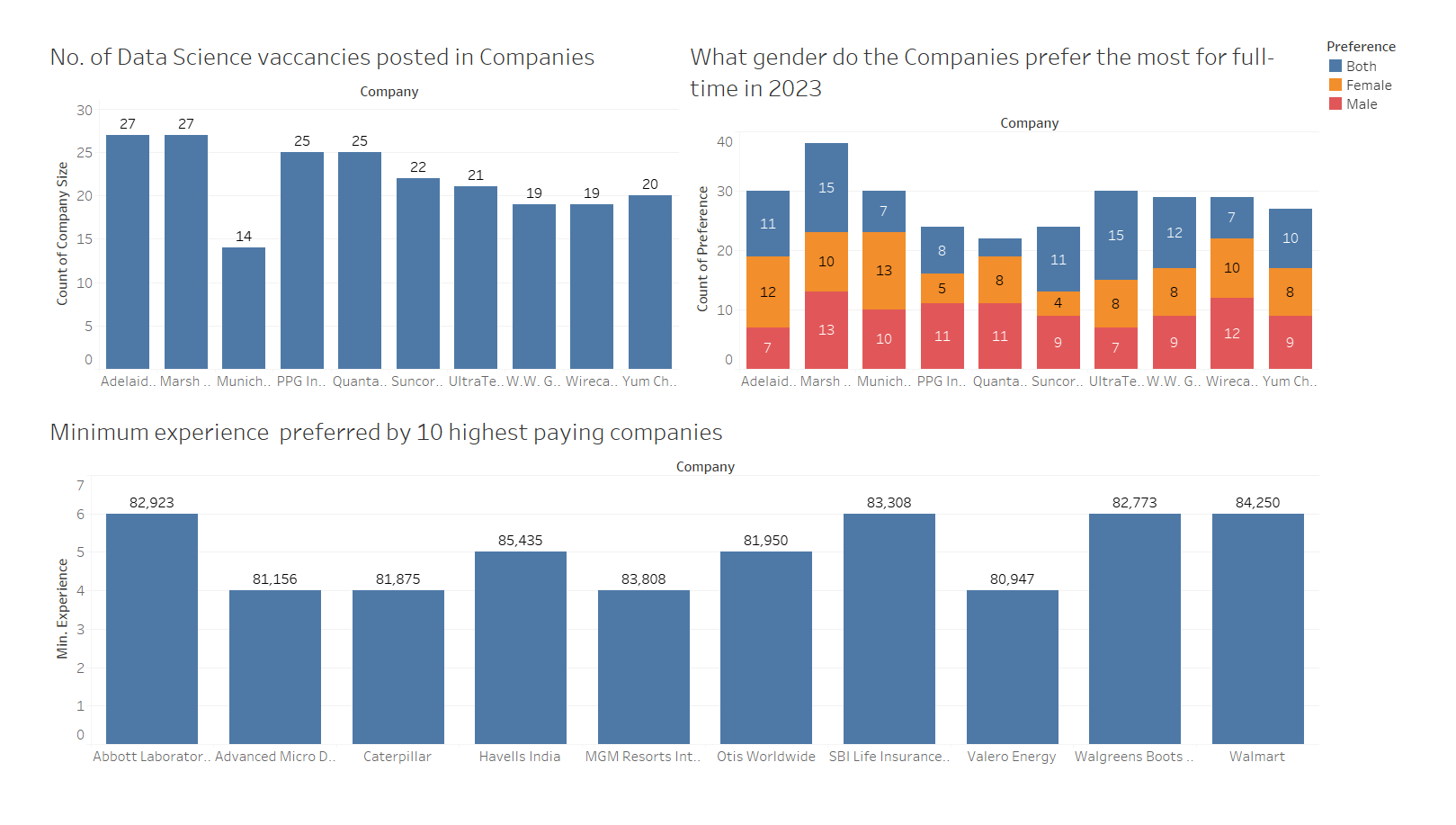
* There is a lack of strong correlation between numeric features.
* Even after employing logarithmic scaling the corelation between maximum minimum and average experience only changes by 0.01 which is negligible.
* This heatmap shows that there might be a different connection with the features other than a linear connection.
* It also points to the fact that the categorical features may be having more influence on the determination of the target variable than numeric features.

### Visualisation

In order to understand the data more deeply and draw analytical insights visual aid is very important, we used tableau for this purpose. Here is some dashboard s we created:

1.

2.

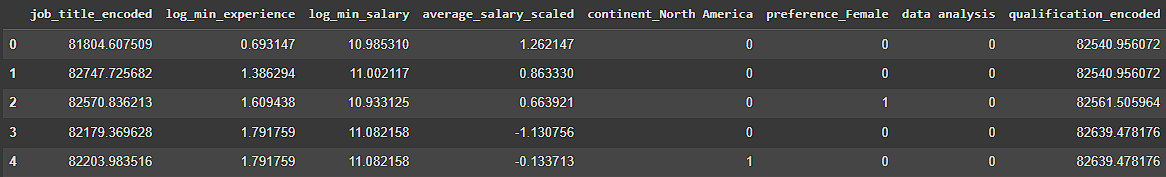
3.

* North America was the continent was the focus of the study, and the main job we looked in to was data scientist.
* The first dashboard talks about the north American comparison. The second one is a more focused comparison on the job title data scientist, and the third one is based on the companies.

### Scaling And Encoding

The total number of features in the data set is 49 after encoding. While using only label and one-hot encoding to all the features according to their characteristics the encoded features would count up to 350 features. The main reason for this change is target encoding which gives a feature its target value and then use it to call upon that value instead of the original one.

This is a head value of some of the features after the encoding:



### Feature Selection

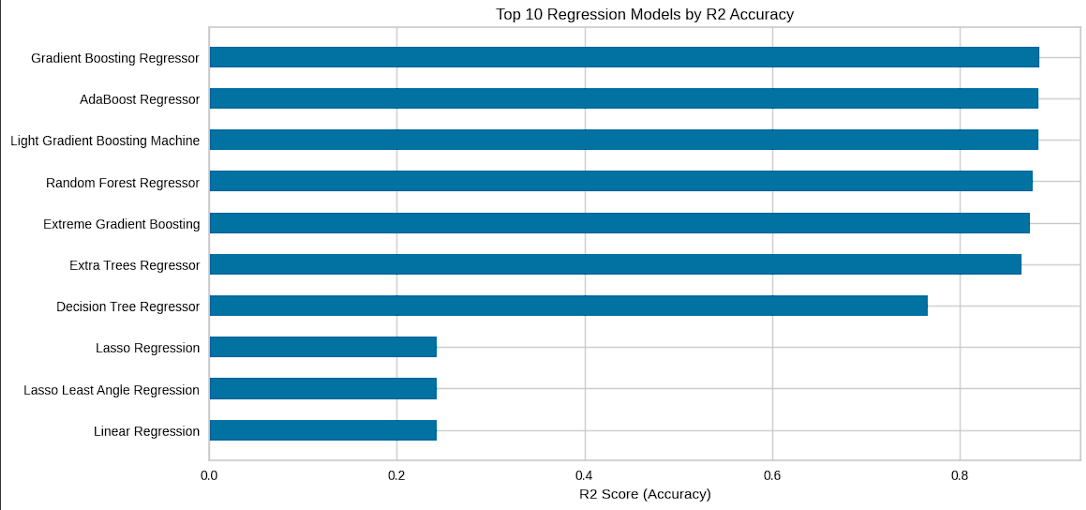
Two methods were employed but, in the end, we moved on with the embedded feature selection the lop features that were selected from this method are:

From this some features were removed in order to create the model such as the features max\_salary, log\_max\_salary were removed.

### Pycaret

The model comparison in pycaret was done using the selected features from the embedded feature selection.

The best models were:



**Gradient Boosting Regressor**

* + **MAE:** 0.2910
  + **R²:** 0.8844
  + **MSE:** 0.1158
  + **RMSE:** 0.3402

**AdaBoost Regressor**

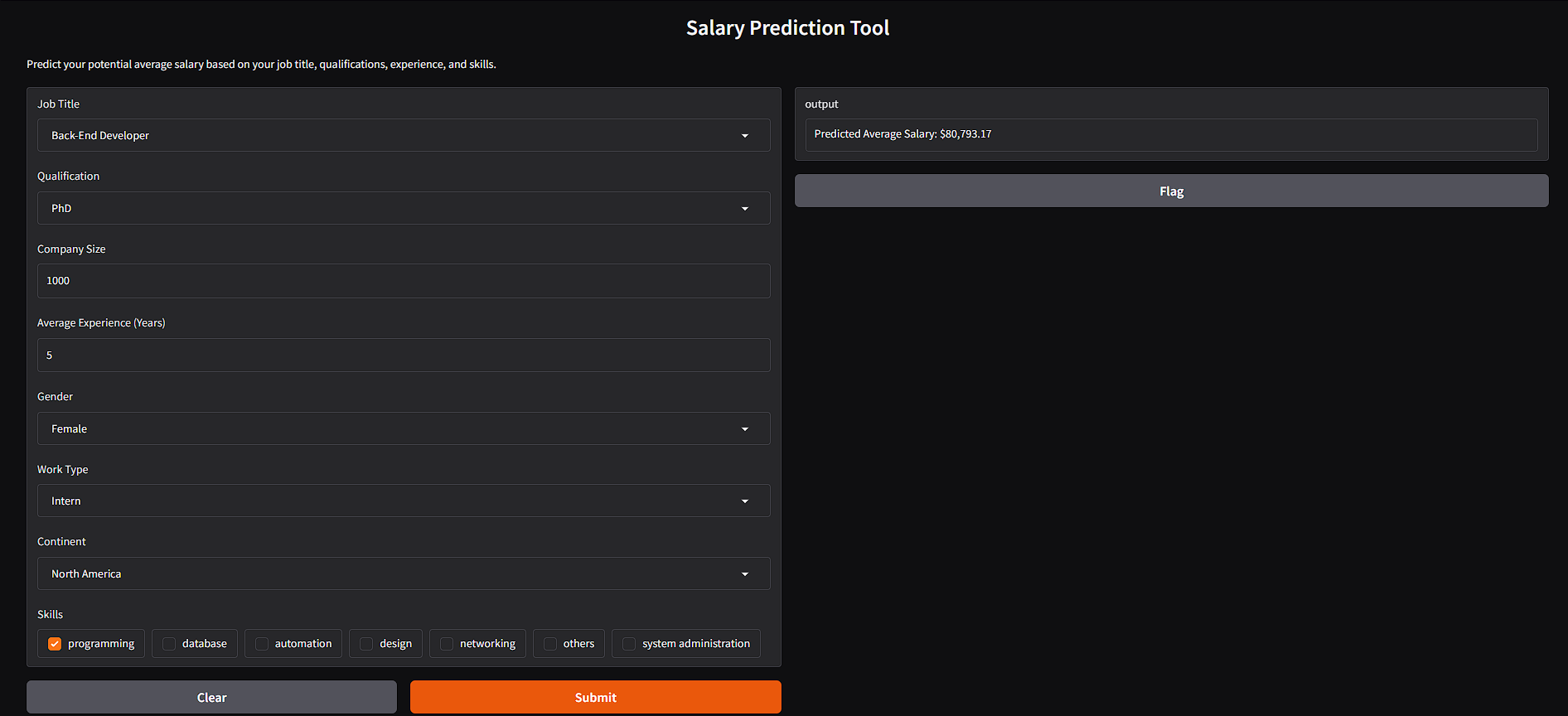
* + **MAE:** 0.2912
  + **R²:** 0.8842
  + **MSE:** 0.1160
  + **RMSE:** 0.3406

### Hyper parameter tunning

Even after the hyper parameter tunning the results barely changed. The accuracy didn’t change while the mean squared error, root mean squared error, mean absolute error values slightly changed (almost at the range of 0.02 for each)

### Deployment

The deployment done using gradio:



The users can input various data including the job title and the qualification they are looking for specifically.

The company size and experience are number based values which are to be given by the user, the category of skills can be used to specify the skill set one might possess.

A gender preference is also mentioned in the deployment.  
Based on these inputs an average salary will be predicted on the right side.

# Discussion

The primary goal of the project was to analyse the job market in the IT sector and predicting the average salary of an IT professional. All these goals have been achieved even though there were many complications and errors in the way it was completed in the end. The most troubling problem was the encoding of the features, and the solution was target encoding which dramatically decreased the dimensionality of the data, from almost 350 features to 49 (350 features were formed after encoding using the normal methods such as label and one-hot). Also mapping skills to a set of skill category was a good option even when considering the deployment. From the start the dataset was very consistent and this had in turn affected the analysis in many ways, this issue was resolved by engineering new features.

# Conclusion

This project has provided great insight toward the changing and rapidly evolving job market in the IT sector, the skill set one should posses as well as the average experience that this field is typically asking for are all a part of the insight gathered from this analysis. Additionally, the salary prediction gives a rough estimation of how much salary one could earn based on the job they are aiming for as well as the skills and experience they possess. As a whole this project has been a lot insightful for an upcoming jobseeker such as ourselves to self-evaluate on ourselves and strive for the betterment of our skills and achieving our goals as an IT professional.

# Contribution

Ajal George (0857357): In the initial phase handled the setting up of a goal and defining the scope, later on helped with the EDA and towards the end contributed in the feature selection part by conducting wrapper feature selection.

Ayush Raj Saxena (0832578): Contributed towards the Exploratory data analysis in the initial phase of the project and later on helped out in the statistical analysis part of the project and helped with implementing pycaret.

Akshay Shanmughan (0851083): Contributed toward finding the dataset and writing documents for the project, toward the end of the project contributed in the feature selection by using embedded feature selection method.

Gautham Salil Panicker (0849202): The first phase of the data cleaning was handled also aiding in the writing of documents for submissions in various phase of the project, towards the end contributed in standardising the data and also helped out with the final project report.

Shreyas Srikrishna Joshi (0837922): Handled the visualisations in the first part and made sure the whole picture on the insights was depicted through the visualisations, also used pycaret to find out which model was best performing.

Vysakh Kavil Padinjarethil Surendran (0857412): From the start kept an eye on the work done by each individual and made any changes that is to be done and implemented the final model which was deduced from the pycaret and deployed the model and also wrote the final project report.

# References

1. <https://pycaret.org/>

# Appendices

<https://github.com/Vysakh01/Capstone>