Salary Prediction System

Data Science - Final Project Presentation

Group Members

Eisha Tir Raazia

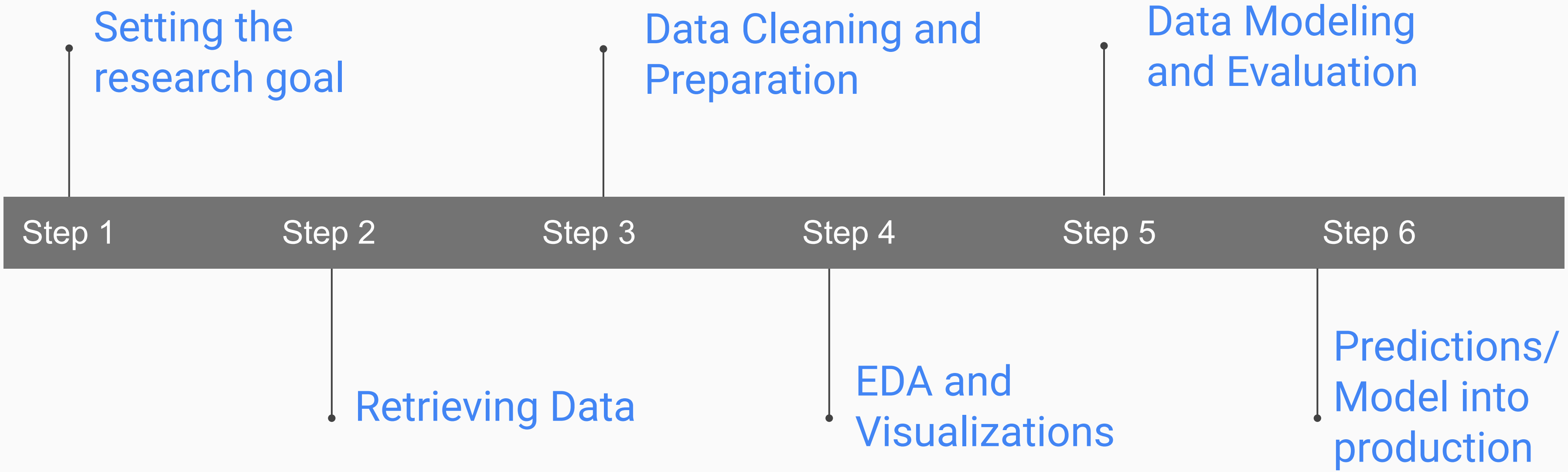
17k-3730

Maham Tariq

17k-3658

Methodology and Steps Involved

Steps Involved

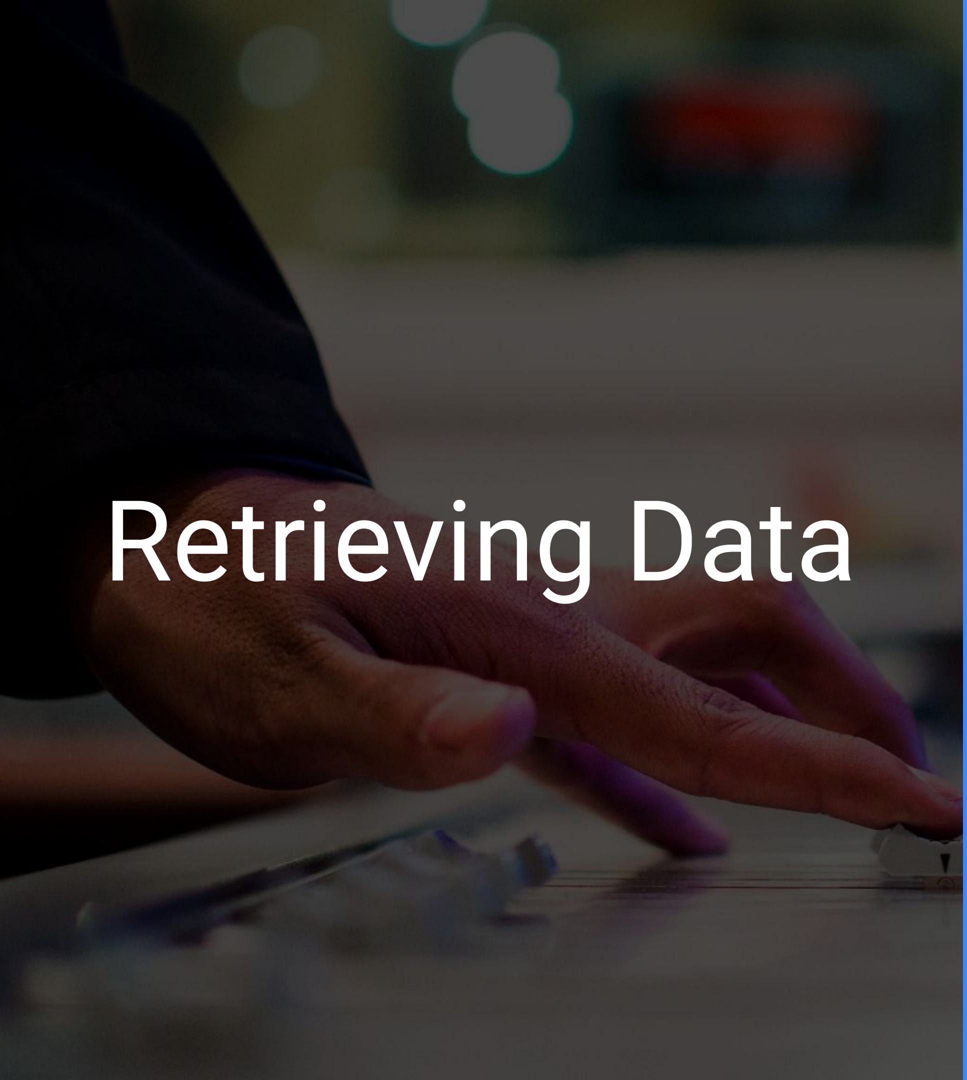


- Step 1 — Setting the research goal
- Step 2 — Retrieving Data
- Step 3 — Data Cleaning and Preparation
- Step 4 — EDA and Visualizations
- Step 5 — Data Modeling and Evaluation
- Step 6 — Predictions/ Model into production

Research Goal

Predict salaries based on different variables

This will help candidate to negotiate their income when they enter the job process.



Retrieving Data

Data Source: Glassdoor.com

Wrote a Selenium based web scraper to extract Companies' data from the site by inputting various job roles.

Challenges we ran into while scraping

- Sign Up dialog pop-up on webpage, at random times that was interrupting Execution.
- Patience for exceptions when browser stops fetching data against a role. (Added max patience of 50)
- Crashes before bulk updating the data into csv.
- A lot of other exceptions to be handled

Data Cleaning & Preparation

After scraping the data, it needed preprocessing and preparation in order to make it was usable for the model.

Data Preparation Steps

We made the following changes and created the following variables:

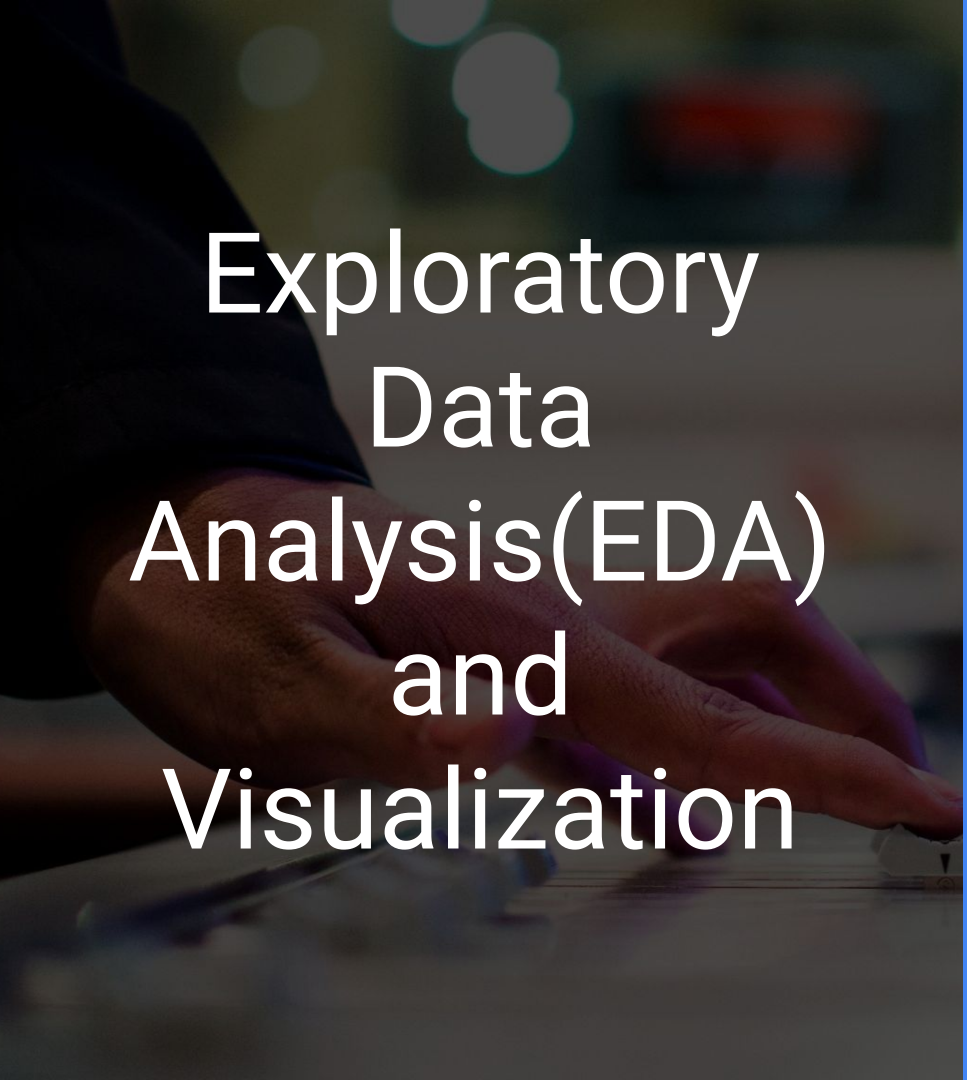
- Removed rows without salary
- Removed Duplicates
- Parsed numeric data out of salary
- Made columns for employer provided salary and hourly wages
- Converted Hourly salary to annual to make column values consistent

Data Preparation Steps

- Made columns for of different skills were listed in the job description:
  - Python
  - R
  - Excel
  - AWS
  - Spark
  - SQL
  - Django
  - Java
  - DB
  - Tableau
  - Azure
  - Hadoop

Data Preparation Steps

- Column for simplified job title(job tag) and Seniority level
- Column for job description(JD) length
- Made a new column for company state
- Transformed founded date into age of company
- Replacing all numeric values with 'Unknown', in ‘Type of ownership’ column



Exploratory Data Analysis(EDA) and Visualization

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Exploratory Data Analysis

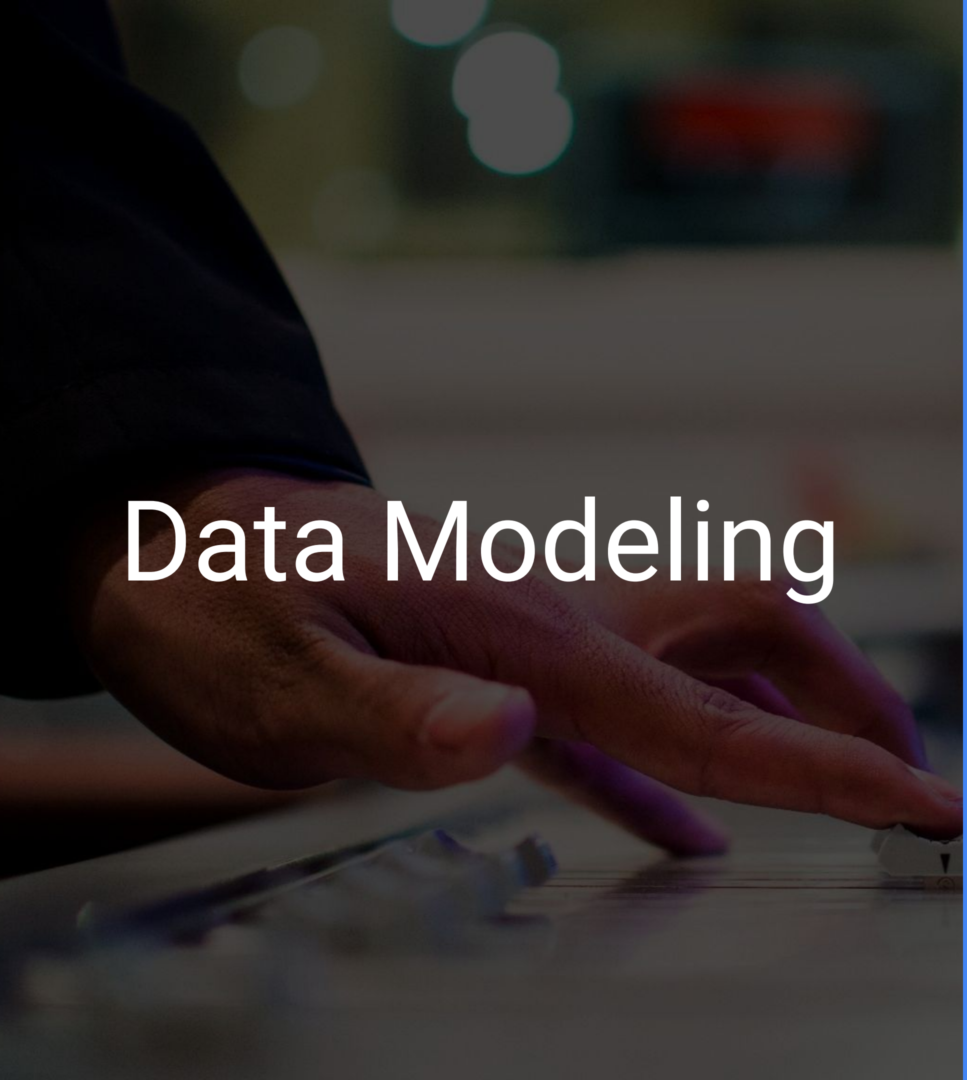
In EDA, we looked at several data distributions and patterns in other to get an idea of how they correlate with the ‘salary’.

Some of the main steps that EDA and visualization part included are:

- Correlation heatmap
- Box plot
- Histogram and data distribution plots

Exploratory Data Analysis

- Scatter Plot with regression lines
- Pivot tables
- Word Cloud image



Data Modeling

We implemented 2 types of models' use cases:

- NN based model to predict Salary using JD embeddings.
- ML regression models to predict salary using other features i.e: company size, job title, etc

Data Modeling ML-based use case

First, we transformed the categorical variables into dummy variables(one-hot data encodings). We also split the data into train and test sets with a test size of 20%.

Data Modeling ML-based use case

We tried three different models:

- Multiple Linear Regression – Baseline for the model
- Lasso Regression – Because of the sparse data from the many categorical variables, We thought a normalized regression like lasso would be effective.
- Random Forest – Again, with the sparsity associated with the data, thought that this would be a good fit.

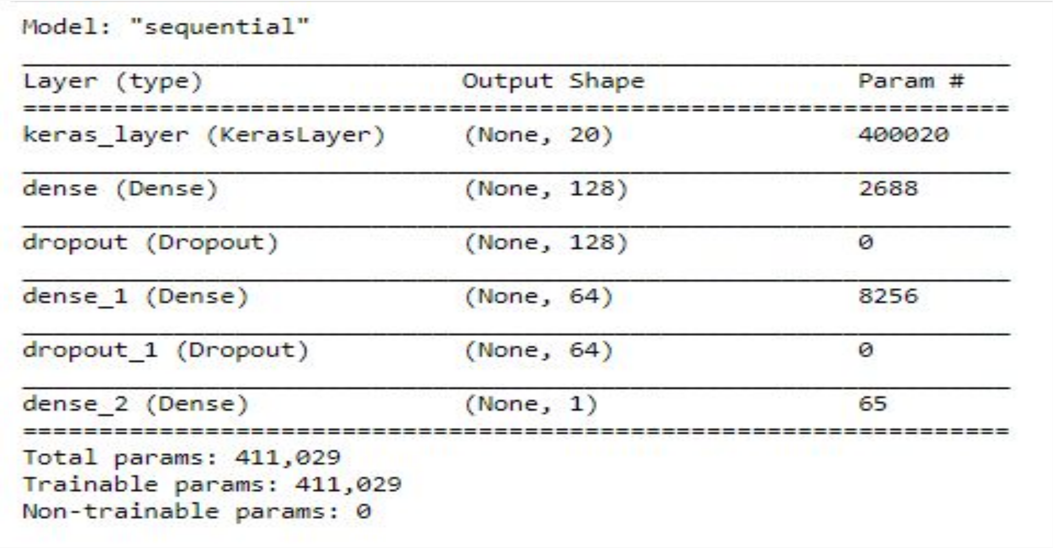
Data Modeling NN and JD embeddings-based use case

Parsed and pre-processed Job description embeddings, and then predicted the salary based through it.

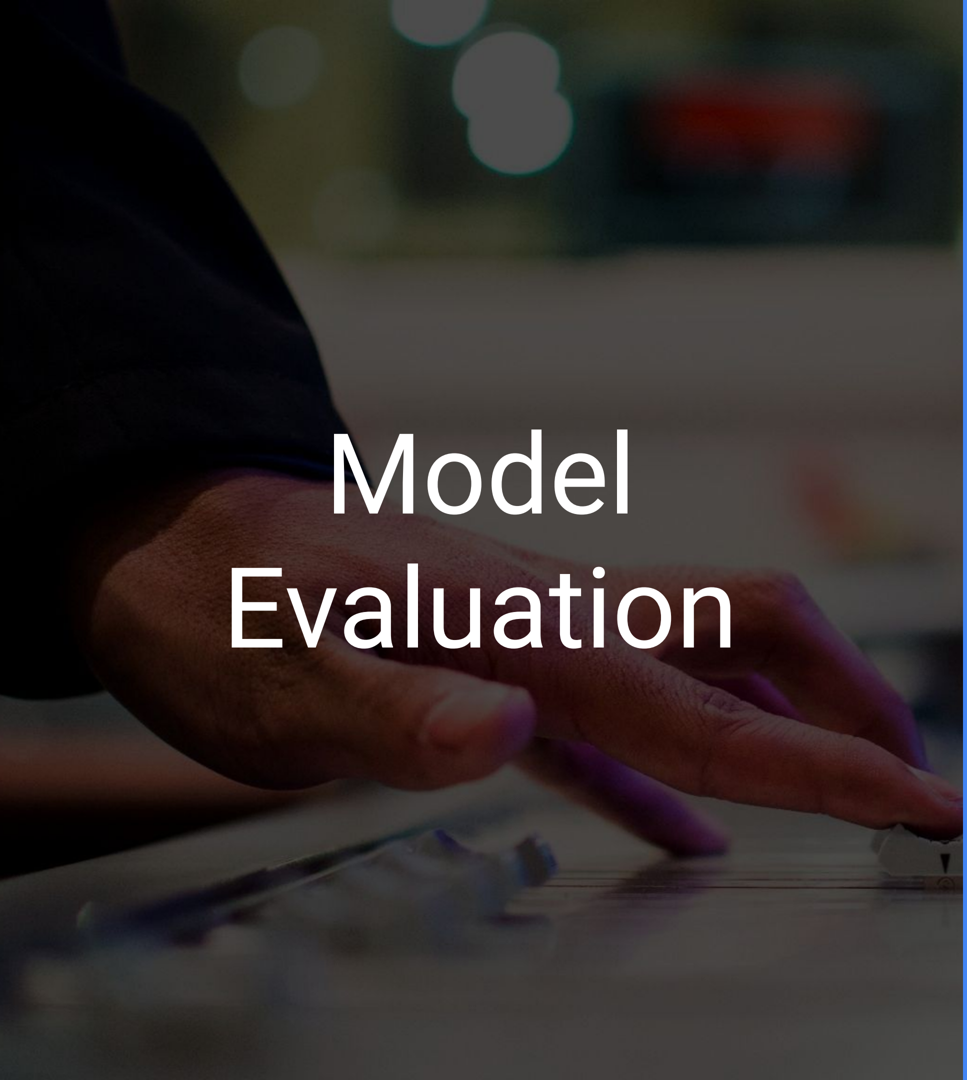
Some of the steps to extract the useful word embeddings included:

- Stop word removal
- Stemming
- Special character and punctuation removal
- Tokenization

Data Modeling NN and JD embeddings-based use case



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
keras_layer (KerasLayer)     (None, 20)                400020
_________________________________________________________________
dense (Dense)                (None, 128)               2688
_________________________________________________________________
dropout (Dropout)            (None, 128)               0
_________________________________________________________________
dense_1 (Dense)              (None, 64)                8256
_________________________________________________________________
dropout_1 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 65
=================================================================
Total params: 411,029
Trainable params: 411,029
Non-trainable params: 0
```



Model Evaluation

We evaluated the three models using Mean Absolute Error (MAE)

We chose MAE because it is relatively easy to interpret and outliers aren’t particularly bad for this type of model.

Model performance

The Random Forest model far outperformed the other approaches on the test and validation sets.

1. Random Forest : MAE = 13.10 (best minimum)
2. Linear Regression: MAE = 01430.09
3. Lasso Regression: MAE = 14.93

Model performance

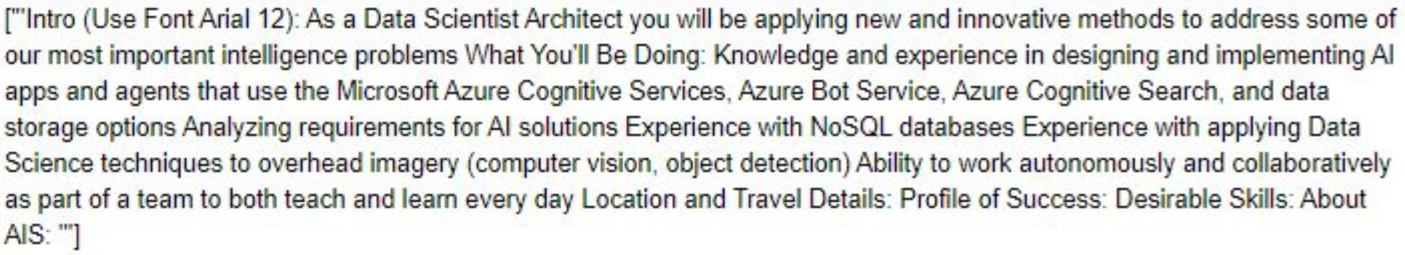
Neural Networks, on JD as input feature, gave the results with minimum errors:

- mae: 11.6152 (best possible minimum value)
- mse: 231.2114
- msle: 0.0337

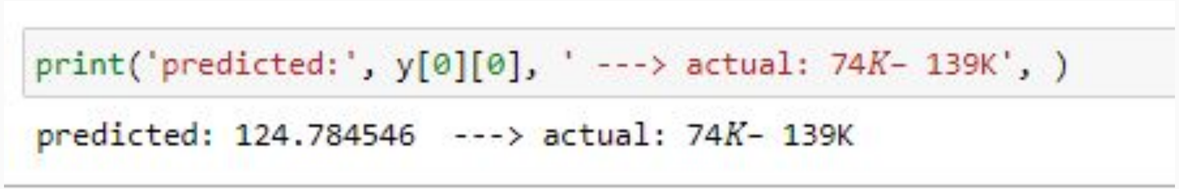
Predictions Endpoint

In order to make predictions, an endpoint microservice through FAST-API is created which returns the predicted salary in the response body.

Predictions - NN-Based Model



["'Intro (Use Font Arial 12): As a Data Scientist Architect you will be applying new and innovative methods to address some of our most important intelligence problems What You'll Be Doing: Knowledge and experience in designing and implementing AI apps and agents that use the Microsoft Azure Cognitive Services, Azure Bot Service, Azure Cognitive Search, and data storage options Analyzing requirements for AI solutions Experience with NoSQL databases Experience with applying Data Science techniques to overhead imagery (computer vision, object detection) Ability to work autonomously and collaboratively as part of a team to both teach and learn every day Location and Travel Details: Profile of Success: Desirable Skills: About AIS: '"]



```
print('predicted:', y[0][0], ' ---> actual: 74K- 139K', )

predicted: 124.784546  ---> actual: 74K- 139K
```

Thank You