Data Collection Lab (094290)

Career Compass: Navigating Salary Insights with AI

Final Project

https://github.com/avishagnevo/LinkedInCareerCompass

Due date: 16.04.2024

Avishag Nevo, ID: 324079763

Almog Ben Simon, ID:206571135

Bana Sadi, ID: 206611477

Omri Lebel, ID: 208772822

1 Introduction

In today's rapidly evolving job market, understanding one's salary relative to others in similar positions is crucial for career development and awareness. While LinkedIn's existing Salary Insights tool[1] provides valuable data, our project introduces a AI-driven feature designed to offer a more personalized analysis of salary wellness. Using advanced machine learning, engineering and statistical techniques, our tool takes into account a broad range of factors —including job role, educational background, experience description, and demographic information— to provide users with tailored insights into where they stand in the salary landscape compared to their peers. Our approach extends beyond numerical comparisons, by employing Large Language Models (LLMs) to generate user-friendly respond that interpret the data into actionable advice. This personalized feedback helps users understand not just where they stand, but also why, and how they might improve their salary. Integrating this feature with LinkedIn's existing salary tool could enrich the platform's utility, delivering insights accessible to all users, regardless of their familiarity with data analysis. Through this project, we aim to provide individuals with the insights needed to make important career decisions, promoting a more transparent job market.

2 Data collection

Our project uses the "/linkedin/people" dataset, scrapped by BrightData. This dataset aggregates professional information from LinkedIn profiles. Within this repository of data, we focus on key columns that are important to our analysis: **position:** The job title of the individual, providing insights into their professional role, free text. **experience_description:** Individual's professional journey, outlining their experiences and achievements, free text. **experience_duration:** Chronicles the timeline of the individual's career, displayed in terms such as "[Jan 2022 - Present (1 year, 9 months), Oct 2020 - Mar 2022 (1 year, 6 months)].", free text. **education_degree:** Reflects the highest level of academic achievement, for instance, "[Master's Degree, Bachelor's Degree].", free text. education_field: Indicates the domain of academic study or specialization, free text. **education_title:** Represents the academic institution from which the individual graduated, partially free text. **experience_length:** Serves as an proxy of the individual's experience. **education_length:** Acts as a measure of the individual's academic level. We used a Python script to identify the most common job titles in the dataset and stored them in a file named "positions_above_300". Since the original data lacks salary information, we've devised a method to estimate salaries, enabling comparisons when users input their own salary details.

"Willingness to share data: Contextual determinants of consumers' decisions to share private data with companies"[2] paper investigates the extent to which peoples' willingness to share private data, including salary or income information, is affected by contextual factors. The authors findings are strongly related for out hypothesis, as they have found that a better intuitive match between the core business a company is engaged in (LinkedIn in our case) and the type of data that is requested (salaries in our case), results in higher proportions of people who are willing to share the corresponding data with the corresponding company. For this Justification we decided collect parameters data for a "smart" synthetic sampling (to be explained later), hence we have scarped the file called "min_max_salary" from a website called "Indeed" to get the minimal, maximal and average salaries for each position found i the most popular positions in the "/linkedin/people" dataset collected in "positions_above_300". we have done the whole scraping process using the Bright Data scraping tool.

In our project, an item is considered to be a single popular position. as we have collected data from "indeed" for all available popular positions, 547 records to be exact (the overlap between the "indeed" available data and the popular positions). The number of popular positions (positions appearing more than 300 timed in the original dataset) were 627, I.e. $\alpha \approx 0.874 = \frac{547}{627}$ which was totally satisfying for us. 5

3 Data Analysis

3.1 Data Analysis

For data cleaning, we've decided to remove entries like '-', '–', etc., from the 'position' column because we wanted to use only positions with valid job titles. This was crucial because the job title is directly related to salary (utilizing the "indeed" data to sample from "families") , having clean data helped with reducing noise in the predictions later on, specially with the position's family finding. Regarding missing values, we have eliminated rows with missing data, particularly in key columns like 'experience' and 'education', as we had enough information (big enough dataset) so we had the legitimatise to let the model train on complete and reliable records only. For fields like 'recommendations_count', we have set defaults where data is missing helps maintain the dataset's size and diversity, which might otherwise be reduced if too many rows were dropped (as we don't know to missing data mechanism). Moving to data transformations, regarding experience and education fields extraction and parsing, we have extracted data from 'experience_description' and 'education' fields to quantify the best proxy for the professional background of the people. For example, converting the 'experience_duration' from text to a numerical representation of total experience in months allows the model to quantify experience in a way that's directly comparable across all people remained in the dataset. We have also used feature engineering to create 'experience_length' and 'education_length', as these features act as proxies for the amount of professional experience and the level of education, respectively. They are likely to be strong predictors of salary because generally, more experience and higher education levels correlate with higher salaries (as we try to create the most realistic dataset we would use them later on to synthesis expected saleries), we took inspiration from the paper "Salary Determinants for Higher Institutions of Learning"[3] as it discusses the determinants of salary for higher institutions of learning. The authors found that level of education and work experience show significantly positive effects on mean salary. 2

3.2 Data Synthesis

In the analysis, we began by preparing the salary data, cleaning it up by removing commas and converting strings to floats to ensure numerical consistency for calculations.

To model the distribution of salaries, we first explored using a Pareto distribution, known for its fit for salary distributions, as a heavy-tail distribution where most people earn less and a small number earn significantly more. Mathematically, a Pareto distribution for a random variable X is defined by:

$$P(X > x) = \left(\frac{x_{min}}{x}\right)^{\alpha} \text{ for } x \geq x_{min}$$

where x_{min} is the minimum possible value of X, and α is a positive parameter. We have reversed engineered alpha such the averaged salary (by "indeed") equal to the expected value each position family, we was inspirited by the Central Limit Theorem idea.

We have also tried using a triangular distribution, which is suited for data with known minimum and maximum values and a peak value (maximal value of the mass distribution function). The triangular distribution is defined by three parameters: a lower limit x_{min}, an upper limit x_{max}, and a mode x_{avg}, where the density function $f(x)$ is given by:

$$f(x) = \begin{cases} \frac{2(x-x_{min})}{(x_{max}-x_{min})(x_{avg}-x_{min})} & \text{for } x_{min} \leq x < x_{avg}, \\ \frac{2(x_{max}-x)}{(x_{max}-x_{min})(x_{max}-x_{avg})} & \text{for } x_{avg} \leq x \leq x_{max}, \\ 0 & \text{otherwise.} \end{cases}$$

This distribution models the salaries as it heavily accounts for the realistic bounds (minimum, maximum) and most likely salary (the mode being the average salary). To further refine the salary estimates, we have adjusted the synthesized salaries by integrating factors such as education and experience, which are known to impact realistic salary levels. We used a custom function to enhance the salary based on a calculated proportion of the difference between the maximum and synthesized salary values, proportionally distributed according to the years of experience and level of education. We took inspiration from the paper[4] discussing the use of the Pareto distribution to estimate earnings, and in particular top earnings, for best imputing top-coded earnings and not systematically understate them. Another paper[5] discusses the properties and usefulness of the Triangular distribution, and argues that it is a suitable distribution to employ in many simulation situations. 4

3.3 Feature Selection

First of all we prepared the data for the feature selection, focusing on "City" and "Recommendations", we filtered cities with representations fewer than 500 to keep data relevance and reviewed the distribution of recommendation counts to initially understand their impact, another thing we did is to simplify the 'languages' feature from a list to a count, to make it a numerical input reflecting multilingual differences which could influence salary, as we only wanted to know if there's a difference the numerical representation was satisfying for us as language is a discrete variable. Using the concatenated condensed text fields education and experience descriptions strings are used as well. After we have prepared the data we have chosen Random Forest Regressor for its robustness against overfitting and capability to handle non-linear data (we know that the data is not linearly distributed), it aids in understanding feature importance, we evaluated the results with RMSE to assess prediction accuracy, and used it for the selection.

Numeric features: - position, followers, education_length, total_experience_years, number_of_languages

Textual features: recommendation_count, experience_description, education

3.4 Visualizations

While doing all this extensive process we have checked ourselves by visualizing the dataset balance and the success in the parsing and transformation process mentioned above, all this can be found in the Appendix. 6

4 AI methodologies

Embedding 3 We aimed to categorize the "position" feature from our LinkedIn dataset into distinct groups to better predict salaries (more standardized and analytically robust dataset). This involved transforming free-text (continues) job titles into discrete categories, later used by the model. This process was established by a Spark NLP pipeline that includes document assembly, tokenization, embedding and vectorization. To prepare the data for this task, we extracted and tokenized position titles from our dataset, identifying the most frequent terms to understand common roles. To generate the embedding we tried two well known encoders, the first is Word2Vec used to capture syntactic and semantic relationships based on local word contexts statistics it has trained on in train time, the other is BERT for its contextual understanding capabilities, as we try to extract some deep insights from the semantic relationships between the words in the job titles. We also condensed text fields like education and experience descriptions into concatenated strings and applied BERT embeddings with Spark NLP to transform the text into vectors that hopefully capture the semantic hints for predicting professional expertise and salary used later on in the feature selection process, salary synthesis and feeded into the model as well.

1NN 3 Utillizing the prepared embedding of the most frequent positions we used a 1-Nearest Neighbor approach to classify raw positions into preselected "position families" utilizing it's robustness for handling complex data categorizations.

Predictive Models we've tried Random Forest Regressor, Linear reggression and XGBoost for the salary prediction. The models were trained on 80% data split, and was evaluated using MSE matric on the rest 20%.

LLM we have utillized google.generativeai library and used GenerativeModel('gemini-pro'), The user enters his current salary, the model extract the processed profile information, we fit the data into the prompt 7a, and show the textual response 7b to the user.

5 Evaluation and Results

In the initial data preprocessing part, the approach we have chosen has left us with more generalizable and robust data for model training (any of them) later on, which was necessary as we heuristically expect the models to work well on real data too. In the data synthesis part, we first evaluated Pareto distribution that was our best card, however, in sampling time, the Pareto model did not fit our data well, producing unrealistic extreme values due to its heavy-tailed nature. We later understood that if we sample from Pareto for each position family we actually make it not Pareto, as Pareto would fit to all families aggregated together. We wanted to make the samples stochastically tailored made, and not only from the aggregated view, we had a little extra research to find the fit, and we found the Triangular distribution which is more suited for the data we have (minimum maximum average). The Triangular approach allowed us to create a realistically distributed salary dataset, providing a statistical foundation for further, well justified analysis. Not surprisingly, the Triangular distribution, aggregated for all the data set, got a shape of Pareto. Regarding classification to "position families" 1NN was very effective and was a good choice for assigning job titles to the nearest category based on their embeddings, and also for its simplicity and explainability. Furthermore we have chosen BERT over Word2Vec after the initial evaluations showed that BERT embeddings were less biased and more effective in capturing the contextual meanings necessary for accurate job title classification using 1NN. Regarding the salary prediction, XGBoost got the lowest MSE on text data (14633.58) w.r.t. Random Forest and Linear Regression (29710.71, 25506.20 respectively). Finally Gimini-pro gave great results as expected.

6 Limitation and reflection

Data Limitations got us removing rows with missing data in key fields, furthermore since salary information was not directly available it had to be synthetically generated. We also had to rely on external sources for salary data (Indeed), and the use of the scraped LinkedIn data concerned us about the consistency and accuracy of the data, specifically regarding the missing data machanism which led us to imply MCAR to ease the process. The limited computational power sometimes necessitated simplifications in our data processing and analysis methods and the project timeline limited our ability to explore more (time consuming) modeling techniques. We had to make modeling choices like using Pareto and triangular distributions for salary synthesis based on theoretical and practical considerations but may not have perfectly captured real-world salary variations. Finally, simplifications in feature engineering (like simplifying language features and filtering out less common data points) might have missed information important for the salary prediction. Our reflection addressing the project's limitations is that we strategically chose to prioritize data quality over quantity by removing rows with missing key fields. This decision likely limited the diversity of our dataset but ensured more reliable analyses and predictions. The absence of salary information necessitated synthetic salary generation using external sources introduce potential biases, as these sources may not fully represent real salary distributions. The limited computational resources and the tight timeline reduced our ability to try more models (potentially better) which also led us to use simpler, theoretically good statistical distributions and assumptions to achieve computational feasibility and avoid overfitting. Overall, these limitations required compromises that probably influenced our project's outcomes by shaping our methodological choices. However, they also fostered a focused and strategic approach to model building within the given constraints.

7 Conclusion

To conclude, we've tried advance the utility of LinkedIn's salary insights by leveraging machine learning techniques to analyze and predict salaries in a more personal manner. We collected, scraped, prepared, and synthesized data, ensuring our model could provide tailored salary evaluations based on the whole profiles, not just job titles. Our data collection involved using scrapped LinkedIn profiles data and enhancing this dataset with scraped salary information from Indeed. This combination allowed us to create a rich dataset despite the absence of salary data on the available LinkedIn dataset. The synthesis process involved generating salary estimates using triangular distribution, which helped us model real-world salary effectively. The feature engineering phase was crucial in refining the predictors which are significant indicators of salary levels. We also applied data transformations to accurately quantify professional backgrounds and skills from raw text descriptions into an informative tailored advise. The challenges we've faced did not stop us providing a tool that could truly benefit LinkedIn users by offering them a clear view of where they stand in the salary landscape compared to their peers.

References

[1] *Introducing "LinkedIn Salary": Unlock Your Earnings Potential*. Available at: https://www.linkedin.com/blog/member/product/introducing-linkedin-salary-unlock-your-earning-potential.

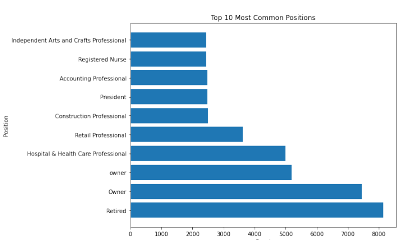
[2] Ackermann, K. A., Burkhalter, L., Mildenberger, T., Frey, M., Bearth, A. *Willingness to share data: Contextual determinants of consumers' decisions to share private data with companies*. Journal of Consumer Behaviour, 21(2), 375-386, 2021. Available at: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cb.2012.

[3] *Salary Determinants for Higher Institutions of Learning in Kenya*. Available at: https://www.academia.edu/73671322/Salary_Determinants_for_Higher_Institutions_of_Learning.

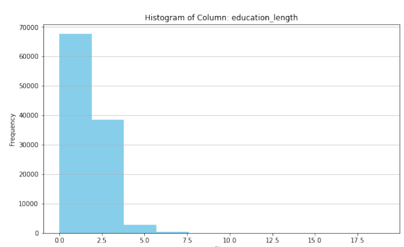
[4] Armour, P., Burkhauser, R. V., Larrimore, J. *Using the Pareto Distribution to Improve Estimates of Top-coded Earnings*. Economic Inquiry, 55(1), 501-524, 2017. Available at: https://www.nber.org/system/files/working_papers/w19846/w19846.pdf.

[5] Fairchild, K. W., Misra, L., Shi, Y. *Using Triangular Distribution for Business and Finance Simulations in Excel*. Journal of Business and Finance, 3(2), 1-15, 2014. Available at: https://www.jstor.org/stable/90001156.

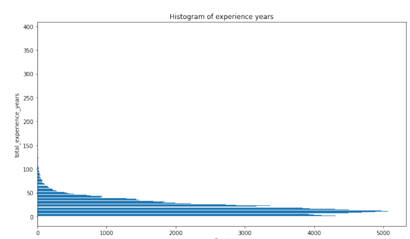
8 Appendix



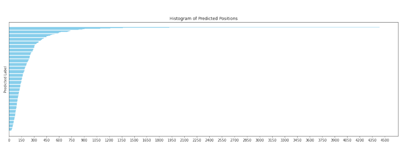
(a) Common Positions



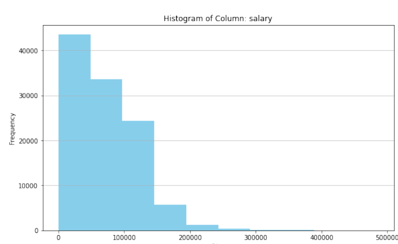
(b) Education Length



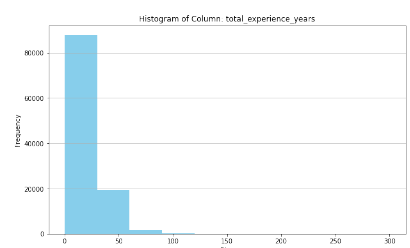
(c) Experience Years



(d) Predicted Positions



(e) Salary



(f) Total Experience Years

Figure 2: Various histograms from the analysis

	position	Predicted_Label
0	interventional radiology tech	service manager
1	system administrator	it specialist
2	structural engineer	it specialist
3	certified project manager	independent consultant
4	account executive, field sales north east, pub...	senior software engineer
5	the counterfeit millionaire... a true crime ad...	retail professional
6	eku student and developing accountant	executive director
7	excellent	retail professional
8	student	sales director
9	corporate sales trainer with extensive leaders...	executive office professional
10	incoming assurance staff	it specialist
11	the sustainable sparky	retail professional
12	wireless system engineer	it specialist
13	realtor	coordinator
14	senior software engineer	it specialist
15	fdm/sla programmer	sales specialist
16	president	coordinator
17	owner/travel consultant	sales manager
18	owner	teacher
19	bachelor of science - bs	senior associate

(a) Word2Vec Nearest Neighbor

	position	Predicted_Label
0	interventional radiology tech	environmental services professional
1	system administrator	system administrator
2	structural engineer	mechanical engineer
3	certified project manager	project engineer
4	account executive, field sales north east, pub...	account executive
5	the counterfeit millionaire... a true crime ad...	warehouse worker
6	eku student and developing accountant	marketing student
7	excellent	accounting
8	student	student-athlete
9	corporate sales trainer with extensive leaders...	territory sales manager
10	incoming assurance staff	quality assurance specialist
11	the sustainable sparky	inside sales representative
12	wireless system engineer	information technology and services professional
13	realtor	realtor
14	senior software engineer	senior software engineer
15	fdm/sla programmer	j.d. candidate
16	president	president
17	owner/travel consultant	owner operator
18	owner	owner
19	bachelor of science - bs	bachelor's degree

(b) BERT Nearest Neighbor

Figure 3: Visualizations of position related tables

position	Low	Average	High
owner	24286.0	59644.0	146481.0
teacher	9.14	19.18	40.25
president	74544.0	136888.0	251374.0
manager	41103.0	76641.0	142906.0
registered nurse	28.99	43.7	65.87
office manager	15.33	22.08	31.79
project manager	56904.0	88410.0	137359.0
administrative as...	13.79	19.14	26.56
ceo	74259.0	145470.0	284969.0
realtor	49476.0	96720.0	189074.0
rn	28.99	43.7	65.87
sales	37788.0	87544.0	202814.0
general manager	35651.0	61407.0	105770.0
attorney	59467.0	104290.0	182900.0
business owner	8.67	18.18	38.14
director	58400.0	106786.0	195262.0
software engineer	75514.0	120112.0	191050.0
vice president	97845.0	158867.0	257947.0
supervisor	11.49	19.29	32.38
operations manager	44211.0	73472.0	122101.0

(a) Dataset Example

position	Low	Average	High	alpha param	pareto_sample
owner	24286.0	59644.0	146481.0	1.6868601	0.22524173626343402
teacher	9.14	19.18	40.25	1.9103585	0.005281236425053848
president	74544.0	136888.0	251374.0	2.1956885	0.211931737198487
manager	41103.0	76641.0	142906.0	2.1565928	0.7165429315346066
registered nurse	28.99	43.7	65.87	2.970768	0.46762529867248737
office manager	15.33	22.08	31.79	3.271111	0.22847315767658727
project manager	56904.0	88410.0	137359.0	2.806132	0.01639676321399719
administrative as...	13.79	19.14	26.56	3.5775704	0.22171992834057552
ceo	74259.0	145470.0	284969.0	2.0428023	0.24383282469915035
realtor	49476.0	96720.0	189074.0	2.047244	3.534943535053907
rn	28.99	43.7	65.87	2.970768	0.07169423749138848
sales	37788.0	87544.0	202814.0	1.7594662	0.4126918387626828
general manager	35651.0	61407.0	105770.0	2.3841822	0.040094527533835356
attorney	59467.0	104290.0	182900.0	2.3267074	0.04532263121833768
business owner	8.67	18.18	38.14	1.9116719	5.294165422499956
director	58400.0	106786.0	195262.0	2.2069607	0.2995497190237779
software engineer	75514.0	120112.0	191050.0	2.693215	0.050289348644250165
vice president	97845.0	158867.0	257947.0	2.6034381	0.31536155519138287
supervisor	11.49	19.29	32.38	2.4730766	0.004401971595697773
operations manager	44211.0	73472.0	122101.0	2.5109189	0.4448724058382354

(b) Pareto Sampled Salary

position	Low	Average	High	tri sample
owner	24286.0	59644.0	146481.0	59690.895
teacher	9.14	19.18	40.25	10.908392
president	74544.0	136888.0	251374.0	136152.58
manager	41103.0	76641.0	142906.0	97039.555
registered nurse	28.99	43.7	65.87	49.696995
office manager	15.33	22.08	31.79	22.760536
project manager	56904.0	88410.0	137359.0	67538.125
administrative as...	13.79	19.14	26.56	19.756634
ceo	74259.0	145470.0	284969.0	147774.06
realtor	49476.0	96720.0	189074.0	164914.61
rn	28.99	43.7	65.87	39.03307
sales	37788.0	87544.0	202814.0	101040.984
general manager	35651.0	61407.0	105770.0	48362.312
attorney	59467.0	104290.0	182900.0	82751.29
business owner	8.67	18.18	38.14	33.960575
director	58400.0	106786.0	195262.0	112850.85
software engineer	75514.0	120112.0	191050.0	100769.234
vice president	97845.0	158867.0	257947.0	169795.84
supervisor	11.49	19.29	32.38	12.816794
operations manager	44211.0	73472.0	122101.0	83328.03

(c) Triangular Sampled Salary

Figure 4: Sampling methods

	position	id	experience_description	experience_duration	education_degree	education_field	education_title	experience_length	education_length
0	Reunification Services, Divorce Coaching, Complex Relationship Counseling, Parenting Coordination, Co-Parenting Education, Adult Mental Behavioral Health,	raklcsw	[I graduated from Jane Addams College of Social Work at the University of Illinois, Chicago (1982) with an MSW. I began as a social worker in a hospital; advanced to clinical specialty areas where I also taught populations from volunteers to pastoral care interns and internal medicine residents; advanced to supervisor of clinical social workers and volunteered at hospice and the American Cancer Society, rising to leadership positions. After developing the hospital case management department, in 1995 I opened my private practice to include consulting and teaching and quickly developed and grew a group practice of clinicians with insurance contracts and employee assistance program (EAP) companies. Administrative duties took me too far away from direct clinical work; returned to my private practice and advanced Critical Incident Stress Management (CISM) training to include specialty areas of aviation and banking industries. I am a frequently requested speaker at conferences and workshops, and have conducted basic and advanced training on a number of clinical issues and topics. Show less, Private Practice Divorce Coach Collaborative Mediation, Adult Treatment of Childhood Trauma, Clinical Social Work, College Faculty; Critical Incident Stress Management Responder, Frequently Requested Speaker. Private Practice Divorce Coach Collaborative Mediation, Adult Treatment of Childhood Trauma, Clinical Social Work, College Faculty; Critical Incident Stress Management Responder, Frequently Requested Speaker.]	[May 1999 - Present 24 years 5 months, May 1999 - Present 24 years 5 months]	[Master's Degree, Bachelor's Degree]	[Clinical/Medical Social Work, Administration, Applied Behavioral Science]	[Jane Addams College of Social Work, University of Illinois, Chicago., George Williams College, Downers Grove, Illinois.]	2	2
1	Bank Teller at Wells Fargo	thomas-jensen-5904b2276	[Complete operational activities while minimizing risk under established policies, educate and connect customers and share the value of mobile banking options, help identify information and services to meet customer financial needs. maintain a secure cashline with high volume cash transactions. Complete operational activities while minimizing risk under established policies, educate and connect customers and share the value of mobile banking options, help identify information and services to meet customer financial needs. maintain a secure cashline with high volume cash transactions., Manage the front-end guest flow; direct and support the service team; educate team members and guests on products and services; and close the books including counting and reconciling cash drawers and daily transactions. Manage the front-end guest flow; direct and support the service team; educate team members and guests on products and services; and close the books including counting and reconciling cash drawers and daily transactions.]	[Jan 2022 - Present 1 year 9 months, Oct 2020 Mar 2022 1 year 6 months]	[]	[]	[]	2	0
2	billing specialist	daily-saez-21aa89a9	[office manager, billing and credentialing specialist office manager, billing and credentialing specialist]	[Feb 2018 - Present 5 years 8 months]	[]	[]	[]	1	0
3	(S.E.R.R.T) Special Education Resource Room Teacher for Prince George's County Public Schools	brandon-m-johnson-97b84231	[Under the direction of the Assistant Supervisor for Instructional Supports, Assessment and Accountability for the Department of Special Education, I monitor, train, and provide technical assistance to schools to successfully implement instructional supports and intervention programs for students with disabilities. I coach educators on the delivery of instruction of identified intervention programs, and collaborate with the principal, general educators and special educators to monitor the fidelity of program implementation and provide recommendations for adjustments if necessary. I consult with The Prince George's County Department of Special Education, Special Education Instructional Specialist and fellow special educators, advising on strategies and activities that may be used to enhance a student's skills through the use of instructional supports and interventions, encourage schools to utilize data to make sound decisions for students and to differentiate instruction, provide modeling and strategies for optimum teaching delivery and student results, develop and provide training to support the knowledge, understanding and delivery of instructional supports and interventions, confer with professional school counselors, school psychologists, social workers and teaching staff to obtain additional testing information and to gain insight on student behavioral disorders affecting learning processes, collaborate with team members to facilitate training on the implementation of instructional supports, interventions and educational assessments, collect data and analyze data to provide feedback to principals and the Department of Special Education for Prince George's County on progress monitoring and fidelity of the implementation of instructional supports and intervention programs, conduct educational assessments, write reports, and provide technical assistance to schools on the administration, scoring and score interpretation of educational assessments. Show less]	[Jul 2021 - Present 2 years 3 months]	[]	[]	[]	1	0
4	Noble Adjusters	thomas-n-webster-0b074248	[Property claims adjuster with extensive background in investigations to include detailed record research, strong personal skills, time management, and reporting skills. Currently hold Florida, Alabama, Mississippi, Louisiana, and Texas adjuster licenses. Xactimate Level 1 certification. Property claims adjuster with extensive background in investigations to include detailed record research, strong personal skills, time management, and reporting skills. Currently hold Florida, Alabama, Mississippi, Louisiana, and Texas adjuster licenses. Xactimate Level 1 certification., Line operations. Line operations.]	[Jun 2021 - Present 2 years 4 months, Dec 2013 May 2021 7 years 6 months]	[]	[]	[]	2	0

(a) Experience and Education Data

	recommendations_count	count
0	NaN	3261293
1	1.0	100406
2	2.0	33011
3	3.0	18378
4	4.0	12116
...	...	...
102	84.0	1
103	124.0	1
104	160.0	1
105	76.0	1
106	99.0	1

107 rows × 2 columns

(b) Recommendations Distribution

current_company:name
Sunnyside Unified...
N/A
null
DFAS
Tokio Marine Insu...
null
null
null
local 242 Laborer...
null
Republic Services
null
null
null
Round Rock ISD
Caregiver Homes
null
Pamela Rak LCSW PC
St. Martin's Press
Flexsteel Industr...

(c) Company Text

title
[Spanish]
[Inglés, Spanish]
[English]
[Spanish]
[Spanish, Chinese]
[English, French]
[Spanish]
[English, Hindi, ...
[Spanish]
[English, Spanish]
[English, Urdu]
[Spanish (DLI eme...
[Tagalog]
[Spanish, English...
[English, Spanish]
[English]
[English, French]
[American Sign La...
[Hindi, Gujarati]
[English]

(d) Languages text

	city	count
0	United States	816231
1	New York, New York, United States	63009
2	Los Angeles, California, United States	39305
3	Houston, Texas, United States	35832
4	San Diego, California, United States	21434
...	...	...
875	Wallingford, Connecticut, United States	503
876	Marina del Rey, California, United States	502
877	Friendswood, Texas, United States	502
878	Fort Walton Beach, Florida, United States	502
879	Greensburg, Pennsylvania, United States	501

880 rows × 2 columns

(e) City Tabular Distribution

education
[]
[]
[]
[硕士 in Applied Data Science from USC Viterbi School of Engineering, 学士 in 信息管理 from 四川大学]
[]
[学士 in 生物化学 from 安徽科技学院, 大专 in 化学生物 from 宿州学院]
[]
[Bachelor of Arts in Early Childhood Education from Harding University, Master of Education in Reading Specialist from Harding University]
[Broker of Real Estate in Real Estate from Arizona School of Business]
[PHD in CPU Architeture from 中科院计算技术研究所]
[汉语国际教育硕士 in 汉语国际教育 from 沈阳师范大学]
[Bachelor of Arts - BA in communication from 美国加州大学圣迭戈分校]
[Master's degree in Civil Engineering from The Johns Hopkins University, 学士 in 土木工程 from 重庆大学]
[Bachelor of Engineering - BE in 电子工程 Electronic&Engineering from 天津大学]
[Certificate in Biotechnology-BioBusiness from 美國喬治城大學, Master of Science - MS in Department of Biological Science and Technology (Org
[M.S.E in Telecommunications from UNIVERSITY OF PENNSYLVANIA, SCHOOL OF ENGINEERING AND APPLIED SCIENCE Philadelphia, B.S in Optical Infor
[Bachelor of Science - BS in Electrical Engineering from George Mason University]
[GED in Music Theory and Composition from Yavapai College]
[Bachelor's degree in International Business/Marketing from University of Cincinnati Carl H. Lindner College of Business]
[VIDEO EDITING / HTML DESIGNS in PASS from Santa Rosa Junior College]

(f) Education Text Example

Figure 5: Dataset text examples

position	salary	total_experience_years	education_length	experience_description	education_field	education_degree	education_title	languages	current_company:name	recommendations_count	
writing and editi...	84499.44	8	0	Description: Sinc...	.	.	.	English Telugu Hindi	TASC (Total Admin...	0	M
construction manager	106991.266	14	0	Administrate cons...	.	.	.	.	Alaska Department...	0	N
crna	232214.73	14	0	经理 经理	.	.	.	.	CRISTAR ENTERPRIS...	0	Po
clinical pharmacist	57.558514	0	2	1. 评估货币经纪公司的交易偏好2...	Applied Data Scie...	硕士 学士	USC Viterbi Schoo...	.	.	0	Ashburn
medical technologist	37.168015	5	0	echo/stress echo/...	.	.	.	.	JUNG MEDICAL CENTER	0	F
sales manager	114815.195	14	2	Assist national a...	生物化学 化学生物	学士 大专	安徽科技学院 宿州学院	.	Vileda Professional	0	Villa Pa
purchasing	24.428633	36	0	teaching and rese...	.	.	.	.	.	0	A
science teacher	72687.65	12	2	-Implemented hand...	Early Childhood E...	Bachelor of Arts ...	Harding Universit...	.	Cabot School Dist...	0	S
talent acquisitio...	115177.234	6	1	Designed and impl...	Real Estate	Broker of Real Es...	Arizona School of...	.	Century Communiti...	1	
senior software e...	190959.81	12	1	Lead to design CP...	CPU Architeture	PHD	中科院计算技术研究所	.	hisilicon	0	
education managem...	61797.117	8	1	中文沉浸式课堂，完全汉语环境，汉语...	汉语国际教育	汉语国际教育硕士	沈阳师范大学	中文 英文 韩语（入门）	Heritage Elementa...	0	Layton, Utah, Uni
cna	22.017155	15	1	Intern at Chinese...	communication	Bachelor of Arts ...	美国加州大学圣迭戈分校	.	CGTN	0	San I
medical biller	18.44935	0	2	参与济南市"嘉丽公馆"项目建设工程...	Civil Engineering...	Master's degree 学士	The Johns Hopkins...	.	.	0	Baltim
sales	87007.73	20	1	2016.7- 2019.7 Hu...	电子工程 Electronic&E...	Bachelor of Engin...	天津大学	.	.	0	San
sales engineer	91464.0	4	3	• Administrated a...	Biotechnology-Bio...	Certificate Maste...	美國南加州大學 National ...	Chinese (maindari...	Taipei Economic a...	0	Was
software engineer	128409.77	0	2	Designed and buil...	Telecommunication...	M.S.E B.S	UNIVERSITY OF PEN...	中文 英语 日语	.	0	Red
software engineer	104510.086	10	1	• Design and impl...	Electrical Engine...	Bachelor of Scien...	George Mason Univ...	English	Evan Walter | Sof...	3	C
executive chef	78658.05	15	1	- Developing dyna...	Music Theory and ...	GED	Yavapai College	English Spanish F...	King Estate Winery	0	E
cpa	104037.17	5	1	Residential Advis...	International Bus...	Bachelor's degree	University of Cin...	.	University of Cin...	0	C
cosmetologist	20.452787	23	1	Specializing in: ...	PASS	VIDEO EDITING / H...	Santa Rosa Junior...	English	Independant	0	S

(a) Raw Data with Salaries

position	salary	total_experience_years	education_length	experience_description	education_field	education_degree	education_title	languages	current_company:name	recommendations_count	city	followers
[-1.202258586883...	84499.44	8	0	[[-1.426543235778...	[[-1.479686379432...	[[-1.479686379432...	[[-1.479686379432...	[[0.3045257329940...	[[-1.097781896591...	0	[[-0.938459753990...	447
[-2.035408973693...	106991.266	14	0	[[-1.420328259468...	[[-1.479686379432...	[[-1.479686379432...	[[-1.479686379432...	[[-1.479686379432...	[[-2.035255908966...	0	[[-0.491760641336...	4
[-0.898244976997...	232214.73	14	0	[[-1.726600289344...	[[-1.479686379432...	[[-1.479686379432...	[[-1.479686379432...	[[-1.479686379432...	[[-1.618400692939...	0	[[-2.786809444427...	2
[-1.847036004066...	57.558514	0	2	[[-0.748495042324...	[[-0.863679826259...	[[-1.884335637092...	[[-2.784879207611...	[[-1.479686379432...	[[-1.479686379432...	0	[[-1.069785833358...	3
[-2.541324377059...	37.168015	5	0	[[-0.817303299903...	[[-1.479686379432...	[[-1.479686379432...	[[-1.479686379432...	[[-1.479686379432...	[[-2.500580072402...	0	[[-2.096100330352...	5

(b) Final Processed Dataset

Figure 6: Data representation before and after processing for analysis

```
genai.configure(api_key="AIzaSyABV8rVzIo_xjRt2wHaYggoRkPcJdOkTfo")
# Load the models

gemini_model = genai.GenerativeModel('gemini-pro')

text_prompt = (f"Your task is to summarize user features from LinkedIn profile with his input salary and predicted salary based on his features."
               f"Please explain to the user the gap between his input salary(from his features) and the predicted salary from our AI model."
               f"based on the following user features(the education_length is the number of diplomas/degrees): {df_as_text}.
               And based on the predicted salary: {predicted_salary} Explain the choice.")
response = gemini_model.generate_content(text_prompt)
text = response.text.replace('•', '*')
```

(a) User prompt input for generating tailored responses using the Gemini model.

```
**User Features:**

* Position: Assurity Title
* Recommendations: 1
* Followers: 2,000
* Salary: $17.52
* Experience Description: 25 years in the title services industry
* Education: Bachelor of Education (BEd) from the University of South Florida
* Education Length: 2 diplomas/degrees
* Number of Languages: 0

**Predicted Salary:**

* $48,756.45

**Gap Analysis:**

Your input salary of $17.52 is significantly lower than the predicted salary of $48,756.45. This gap can be attributed to several factors:

* **Years of Experience:** Your 25 years of experience in the title services industry is a valuable asset. The predicted salary reflects the higher earning potential associated with this level of experience.
* **Education:** While you have a Bachelor of Education degree, it is not directly related to the title services field. The predicted salary takes into account the importance of field-specific education or specialized certifications.
* **Certifications or Training:** It is not clear from your profile whether you have any certifications or specialized training in the title services industry. Such qualifications can significantly increase your earning potential.
* **Industry Demand:** The title services industry may be experiencing increased demand for skilled professionals, leading to higher salaries.
* **Location:** The geographical location of your work can also influence salary expectations. The predicted salary may reflect higher earning potential in the area where you reside.
...
* Obtain additional certifications or specialized training in the title services industry.
* Explore opportunities for professional development and continuing education.
* Network with industry professionals and stay informed about industry trends.
* Consider relocating to an area where salaries for your skills are typically higher.
```

(b) Tailored response for user salary gap analysis generated by the Gemini model.

Figure 7: Illustrations of the Gemini AI model's user interaction