CSE 6242
Data & Visual Analytics
Final Report

Fangzhe GU, Junzhe LIAO,
Shiyu JI, Xi WANG

1 INTRODUCTION

This study is designed to provide a comprehensive analysis of the employment landscape for science and engineering graduates across major Chinese cities such as Beijing, Shanghai, Guangzhou, and Shenzhen. According to (Tang, 2020), there is a significant shortage of skilled professionals in many sectors, particularly in technology-driven fields. This gap highlights the importance of understanding the range of career paths available to graduates, from continued academic research, as detailed by (Shui and Hu, 2022), to roles within governmental organizations, which are increasingly popular according to (Tao, Xie, and Wang, 2023). Additionally, there is a notable rise in entrepreneurial ventures among graduates, as explored in (Wang et al., 2021).

Our research intends to dissect employment statistics meticulously, taking into account educational backgrounds, key competencies, and prevailing industry trends. By doing so, we aim to provide a detailed perspective on the professional avenues open to these graduates. This approach not only aids in bridging the existing skill gap but also supports graduates in making informed career choices based on robust data.

Despite the inherent limitations in the available data, our methodology will leverage sophisticated predictive models to ensure accuracy and reliability. Following the guidelines set forth by (EY, 2021), we emphasize the use of data analytics in managing and understanding risks associated with employment trends. By employing these methodologies, our study strives to offer invaluable insights that will aid graduates in navigating their future careers in a more informed and strategic manner.

2 METHOD

2.1 Exploratory Data Analysis

2.1.1 *Data Cleaning*

In the initial phase of our research, we undertook a meticulous data cleaning process utilizing OpenRefine (Verborgh and De Wilde, 2013), a powerful tool for handling and rectifying disorganized datasets. The primary step involved the organization and refinement of the raw data, which exhibited substantial irregularities and discrepancies initially. Our categorization protocol was based on several pivotal attributes including geographic location, levels of education, and extent of job experience.

The data was systematically divided into various districts corresponding to their geographic locations, thereby facilitating region-specific analyses. Educational qualifications were harmonized and classified into distinct categories: college, bachelor's, master's, and doctoral degrees. This standardization was crucial for maintaining consistency across the dataset. Similarly, job experience was categorized into definitive ranges: 1-3 years, 3-5 years, 5-10 years, and over 10 years, allowing for a more nuanced analysis of career progression and opportunities.

Furthermore, we addressed typical data integrity issues such as missing entries, duplicate records, inconsistent formatting, and potential inaccuracies. Our approach ensured the precision and consistency essential for robust analytical outcomes. One of the more challenging aspects of this phase involved resolving overlapping salary ranges. To manage this, salary data were first averaged and subsequently segmented into six distinct ranges. These comprised a segment under 10 thousand, a segment over 30 thousand, and four intermediate segments, each spanning a 5,000 interval within the 10 thousand to 30 thousand range.

In addition to these measures, we standardized variables pertaining to company types and sizes, significantly enhancing the dataset's clarity and applicability for subsequent analyses. This comprehensive organization and standardization process not only streamlined the dataset but also established a solid foundation for the intricate research and analysis that followed. The meticulous attention to data quality and integrity in this phase was instrumental in ensuring that subsequent findings were based on reliable and accurate data, thus providing valuable insights into the employment landscape for science and engineering graduates in major Chinese cities.

2.1.2 *Analysis of Salary & Company Type*

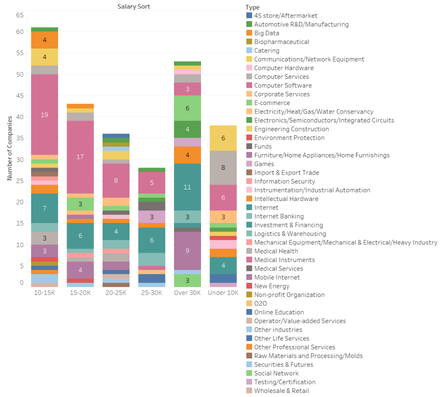


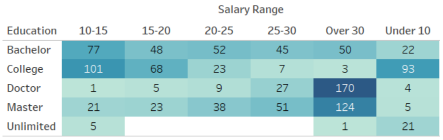
Figure 1—Salary-Company Type Stacked Bar Chart in Shanghai

This stacked bar chart 1 statistically illustrates the relationship between the number of companies and salary distribution in different industries. The picture shows that the highest bar is 10-15K, which may mean that more positions in the market are entry-level or mid-level. As salary grades increase, the number of companies decreases, indicating that high-paying jobs are scarce. However, the total number of companies with the highest salary range (Over 30K) is the second-highest bar in the figure, second only to 10-15K. This may be because the classification of this part is not detailed enough.

In the higher salary range (Over 30K), the number of companies in certain industries such as e-commerce and semiconductors is significantly higher, which may indicate that specific positions in these fields have higher salaries. In addition, technology-related industries (such as computer software, Internet, Mobile Internet) are widely distributed in multiple salary ranges, showing the breadth and diversity of their market needs.

We could use hypothesis testing such as the chi-square test to determine whether these observations are statistically significant. Additionally, estimates of average salary levels and measures of variability in salary ranges across industries would further enrich this analysis. In short, this chart provides a visual understanding of the job market and salary distribution across different industries.

2.1.3 *Analysis of Education & Salary Range*



	Salary Range					
Education	10-15	15-20	20-25	25-30	Over 30	Under 10
Bachelor	77	48	52	45	50	22
College	101	68	23	7	3	93
Doctor	1	5	9	27	170	4
Master	21	23	38	51	124	5
Unlimited	5				1	21

Figure 2—Education-Salary Range Highlight Tables in Shenzhen

The Education-Salary Range Highlight Tables for Shenzhen vividly illustrate the correlation between educational attainment and salary distribution among graduates. The data shows a clear progression in earning potential correlated with higher educational levels.

Individuals with college degrees are predominantly found in the lowest salary bracket, typically earning under 10,000. This highlights the limited economic prospects for those with lower educational qualifications. In contrast, bachelor's degree holders most commonly earn between 10,000 to 15,000, indicating a slight improvement in salary with increased education.

More significantly, those with master's and doctoral degrees tend to earn above 20,000, with doctoral degree holders frequently appearing in the highest salary bracket of over 30,000. This trend underscores the substantial economic benefits that advanced degrees confer, aligning higher education with higher pay.

The mid-range salary brackets (15,000 to 25,000) are most commonly associated with bachelor's and master's degree recipients, illustrating a moderate but noticeable increase in earning potential with each additional level of academic achievement.

These findings from Shenzhen underscore the importance of higher education in enhancing economic outcomes. They offer valuable insights for policymakers and educational institutions aiming to better align educational offerings with the economic opportunities of the market, thereby optimizing career prospects for graduates.

2.1.4 *Analysis of Job Seeker's Experience*

We also collect data about job seekers' experience in megacities, here we take Shanghai as an example for an in-depth analysis. As the pie chart 3 shows, only a relatively small portion of senior-level positions require more than 10 years of experience, indicating that there are fewer senior-level positions or that there is a high barrier to entry for these roles. And mid-level positions (1-3 years and 3-5 years) together make up a significant portion of the job market, reflecting a robust demand for professionals who have developed their skills and can contribute effectively to the workforce. For advanced Mid-level positions (5-10 years), there shows a modest demand for positions requiring this range of experience, possibly pointing to a more niche segment within the job market that values seasoned expertise. And with regard to the no-experience-required entry-level positions, which is the most significant segment, showcasing a job market ripe with opportunities for new graduates, individuals starting their careers, or those looking to switch industries without prior experience in the new field.

The data suggests that Shanghai's job market is particularly welcoming to entry-level candidates, while still offering substantial opportunities for those with moderate experience levels. However, highly experienced professionals may face stiffer competition for a smaller pool of available positions.

2.1.5 *Analysis of Job Seeker's Educational Background*

Educational backgrounds play an important role in the distribution of salaries. Take the employment situation in Shenzhen as an example. The radar chart in figure **??** presents salary ranges across different educational levels in Shenzhen. The farther the line stretches, the higher the proportion of the degree in the corresponding salary range. Also, the stacked area chart in figure **??** depicts the salary distributions across different educational levels in Shenzhen. Taking the two figures into account, we can conclude that PhD holders have the highest percentage in the salary range above 30K, while individuals with a college degree have a higher percentage in the lower salary range (e.g., under 10K). Holders of bachelor's and master's degrees are more common in the mid-salary range, such as 15-20K and 20-25K. Meanwhile, the prevalence of associate degrees increases in the lower salary brackets and decreases sharply as salary increases.

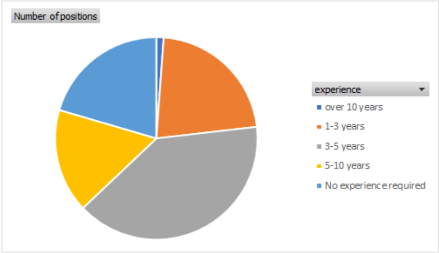


Figure 3—Experience-Salary Pie Chart in Shanghai

Overall, these charts reinforce the notion that there is a correlation between the level of education and the salary range, and higher educational attainment can be associated with higher salaries, with the most significant distinction observed at the highest salary bracket.

2.2 Model Establishment

In the current phase of our project, we are focusing on developing an XGBoost predictive model to estimate the employment salaries of young engineering students. The core of our predictive analysis is the XGBoost (Extreme Gradient Boosting) (Chen and Guestrin, 2016) algorithm, a decision tree-based ensemble machine learning algorithm that utilizes a gradient boosting framework. While deep learning models often represent the state-of-the-art in many predictive tasks, XGBoost was specifically chosen for its comparative simplicity and rapid training capabilities. Despite its simpler structure, XGBoost competes favorably in performance, particularly due to its efficiency, flexibility, and capability to handle sparse data, making it an appropriate choice for our diverse, multi-dimensional dataset. Model training is conducted using the Python-based XGBoost library.

To integrate XGBoost into our analysis, we first perform comprehensive preprocessing on our dataset (He et al., 2008), which involves encoding categorical variables into a machine-readable format, estimating missing values for continuous features (Kuhn and Johnson, 2019), and scaling numerical features to bring them into a comparable range. Our dataset incorporates quantitative academic metrics, qualitative skill assessments, and demographic information, which we have intelligently encoded to maintain information integrity. After preprocessing, we proceed to model training.

One of the key advantages of XGBoost over more complex models is its exceptional handling of imbalanced datasets; it automatically assigns a higher weight to the minority class, thus compensating for any potential biases. This feature, combined with its robustness and shorter training times, makes XGBoost an efficient tool for rapidly deploying scalable predictive solutions. We have divided our dataset into a training set and a test set in a 70:30 ratio to facilitate learning and subsequent evaluation of the model's predictive capability.

On the backend integration front, we are implementing a RESTful API using Flask, which will serve the trained model. This lightweight web service will handle incoming HTTP requests, extract feature data from the request payload, and respond with prediction results. Our goal is for the backend to seamlessly interface with the front end, providing a smooth user experience for stakeholders to interact with our predictive tool via a web interface. The deployment of this API will be containerized to enhance scalability and facilitate updates as the model evolves (Matthias and Kane, 2015).

3 VISUALIZATION

3.1 Front-end Platform

1. **Front-end Architecture:** The client-side is built using *React.js*, which provides a robust framework for developing user interfaces. The project utilizes *Vite* as a build tool, which significantly improves the development startup time and provides instant module reloading. For routing and state management, *React Router* and *Redux* are employed, respectively, ensuring seamless navigation and consistent state across the application.
2. **User Interface Design:** The UI is designed with interactivity in mind, utilizing *Recharts* and *Echarts* for responsive and visually appealing data visualizations. *Material UI* is integrated to craft a modern look and feel, enhancing the overall user engagement through aesthetically pleasing components.
3. **Back-end Configuration:** The server-side is powered by *Node.js*, which handles the processing of CSV data and facilitates the interaction between the front-end and back-end. It includes functionalities for filtering data based on user inputs, allowing for customized data presentations (more details about presentation, please see Section 2.2 on model establishment). Furthermore, the back-end incorporates predictive models that enable users to manually select predictions for salaries, adding a layer of interactivity and personalized experience.
4. **API Testing:** The APIs are thoroughly tested using *Postman*, ensuring reliability and performance of the data interactions within the application.

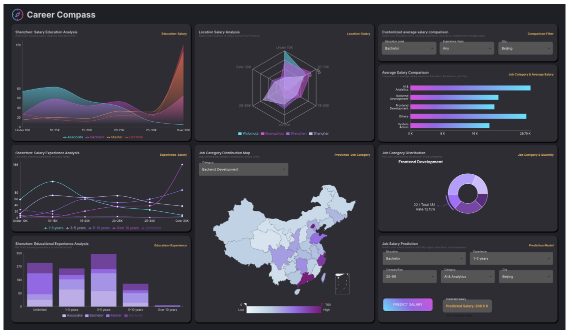
3.2 User Interface

The Career Compass project features a meticulously designed dashboard (Fig 4) that leverages a grid layout to distinctly categorize and display various data sections. This layout adapts seamlessly across different device screens, ensuring a consistent user experience. The interface is highly interactive, allowing users to manually select specific job information, which aids them in clarifying their career direction.

3.2.1 *First Column*

This column visually represents the interrelationships among salary, education level, and working experience of Shenzhen through three distinct types of charts. User can view the corresponding data by placing the mouse in different positions of the charts.

1. **Stacked Area Chart:** This chart illustrates the disparities in the number of people within different salary ranges across various educational levels.(Fig.5)
2. **Line Chart:**Depicts the relationship between salary and work experience, highlighting trends over various experience levels.(Fig.6)
3. **Bar Chart:** Shows the correlation between work experience and educational levels, providing a clear comparison of employment statistics.(Fig.7)



Education Level: Bachelor
Experience Years: Any
City: Beijing

Category: Backend Development

Education: Bachelor
Experience: 1-3 years
CompanySize: 20-99
Category: AI & Analytics
City: Beijing
Predicted Salary: Predicted Salary: 206.5 K

Figure 4—Dashboard of Career Compass

3.2.2 *Second Column*

This column is dedicated to displaying data by geographic region, enhancing understanding of regional variations and trends.

1. **Radar Chart:** This chart displays salary disparities across four cities: Urumqi, Guangzhou, Shenzhen, and Shanghai. Users can interact with the chart by clicking on the legend to select different cities for comparison. (Fig.8 and 9)
2. **China Map:** This map shows the distribution of job categories across various provinces in China. The categories displayed include AI & Analytics, Back-end Development, Front-end Development, System Admin, and Others. Users can select the job category they are interested in via a drop-down menu, allowing for targeted exploration of employment data across the country. (Fig.10 and 11)

3.2.3 *Third Column*

This column is designed to provide users with a customizable data visualization experience, including options to filter data and engage with predictive features.

1. **User Customization Options:** Users can customize views by selecting education level, years of experience, and city. This feature enhances their ability to filter the obtained data and acquire a more refined understanding of career trajectories and industry trends. (Top of Fig.12)
   - **Horizontal Bar Chart:** This chart displays average salaries for five job categories based on user-selected criteria. It visually represents how different factors influence salary scales within specific sectors, helping users make informed decisions based on potential earnings. (Middle of Fig.12)
   - **Pie Chart:** This chart shows the distribution and proportion of job positions based on current user selections. It provides a visual breakdown of job counts by category, highlighting the relative sizes of each sector within the user-defined context. (Bottom of Fig.12)
2. **Prediction Model:** This advanced feature allows users to predict annual salaries by choosing among five different categories and click 'Predict Salary' button.(Fig.13)

4 EVALUATION OF MODEL

Initially, we assessed the model's accuracy using R-square, which provides a measure of how much of the variance in the dependent variable (employment outcomes) can be predicted from the independent variables. While R-square is a straightforward indicator of model performance, it does not always provide the full picture, especially in complex datasets with many variables. To provide a more comprehensive evaluation, we also considered the adjusted R-square, which adjusts the statistic based on the number of predictors in the model, thereby providing a more accurate assessment in the context of multiple regression.

Moreover, to ensure that our model's performance was not a result of overfitting the training data, we employed k-fold cross-validation. This technique involves partitioning the original sample into k equal-size subsamples, using a single subsample as the validation data for testing the model, and the remaining k-1 subsamples as training data. This process is repeated k times, with each of the k subsamples used exactly once as the validation data. The results from the k-folds were then averaged to produce a single estimation, providing us with a more reliable representation of the model's predictive performance on unseen data.

For our XGBoost model, we conducted a parameter grid search to determine the optimal settings. This approach led us to select approximately 1000 weak learners and an alpha value of 0.3, which optimized the balance between model complexity and prediction accuracy. Based on the outcomes of these evaluations, we can confidently iterate on our model to enhance its predictive power, ensuring that it is well-suited to serve the needs of stakeholders in the context of young engineering graduates' employment outcomes.

Model	**Training Set** R^2	**Testing Set** R^2
XGBoost	0.67	0.54
LR	0.15	0.12
SVR	0.40	0.36
NN	0.35	0.32

Table 1—R-square Values for Different Models

5 CONCLUSION & FUTURE PROSPECT

Our project has effectively demonstrated the viability of using the XGBoost algorithm, optimized through a parameter grid search, for predicting employment outcomes for young engineering graduates. Despite its success, the model currently relies on a static dataset from Boss Zhipin, which limits its ability to reflect real-time job market changes and potentially impacts predictive accuracy. Moreover, as the model's sophistication increases, the necessity for a robust database to facilitate better data management and analysis becomes apparent.

To enhance our platform's utility and accuracy, we plan to integrate real-time data feeds, replacing our static dataset with dynamically updated information that mirrors current market conditions. This adjustment will likely improve our model's responsiveness to market fluctuations. Additionally, we aim to refine our predictive models continually by exploring advanced algorithms and incorporating more varied data, such as economic indicators and company-specific details.

Parallel to these technical enhancements, developing a scalable database system will be a priority. This system will support larger data volumes and more complex analyses, essential for both improving backend models and upgrading our data visualization tools. This approach will not only streamline data management but also enrich the user experience by providing more interactive and insightful visual analytics.

By addressing these points, we aim to make our platform a more accurate and user-friendly resource for graduates and employers alike, ensuring it remains a relevant and powerful tool in the employment prediction and analysis landscape.

6 DISTRIBUTION

All team members have contributed a similar amount of effort.

Name	**Effort**
Fangzhe GU	25%
Junzhe LIAO	25%
Shiyu JI	25%
Xi WANG	25%

7 REFERENCES

[1] Chen, Tianqi and Guestrin, Carlos (2016). “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 785–794.

[2] EY (2021). *Role of Data Analytics in Risk Management*. https://www.ey.com/en_in/risk/role-of-data-analytics-in-risk-management. Accessed: 2024-02-25.

[3] He, Haibo, Bai, Yang, Garcia, Edwardo A, and Li, Shutao (2008). “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, pp. 1322–1328.

[4] Kuhn, Max and Johnson, Kjell (2019). *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC.

[5] Matthias, Karl and Kane, Sean P (2015). *Docker: Up & Running: Shipping Reliable Containers in Production*. "O’Reilly Media, Inc."

[6] Shui, Yongsheng and Hu, Yue (2022). “Research on employment guidance strategies for graduate students in science and engineering—taking Xiamen University”. In: *University Logistics Research*.

[7] Tang, Biao (2020). “Opportunities and Challenges for Graduates of Colleges and Universities in Shenzhen City in the Context of the Construction of the Guangdong-Hong Kong-Macao Greater Bay Area”. In: *The 3rd International Conference on Economy, Management and Entrepreneurship (ICOEME 2020)*. Atlantis Press, pp. 407–413.

[8] Tao, Lin, Xie, Yongjiang, and Wang, Yi (2023). “Employment status and countermeasures for undergraduates in science and engineering colleges—taking the School of Cyberspace Security of Beijing University of Posts and Telecommunications”. In: *Beijing Education*.

[9] Verborgh, Ruben and De Wilde, Max (2013). *Using openrefine*. Packt Publishing Ltd.

[10] Wang, Shan, Yue, Jianshe, Wang, Xiaofang, Zhu, Wenting, Ding, Xiao, and Gao, Yihong (2021). “Analysis of the influencing factors of innovation and entrepreneurship education on employment in science and engineering under the background of "Internet"”. In: *Light Industry Science and Technology*.

8 APPENDIX

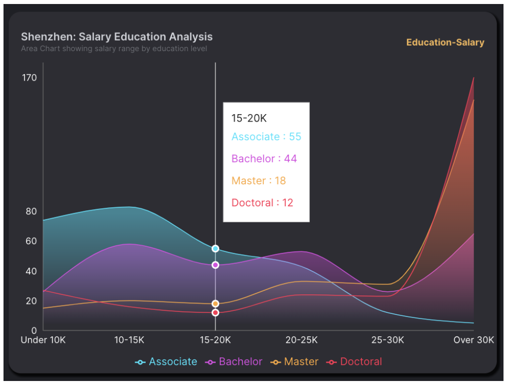


Figure 5—Stacked Area Chart

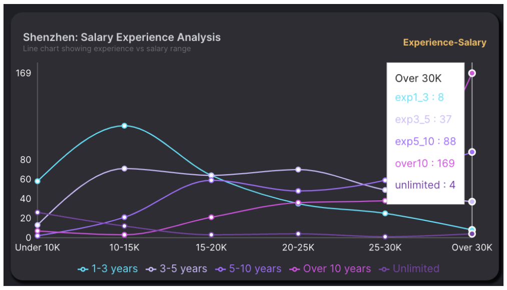


Figure 6—Line Chart

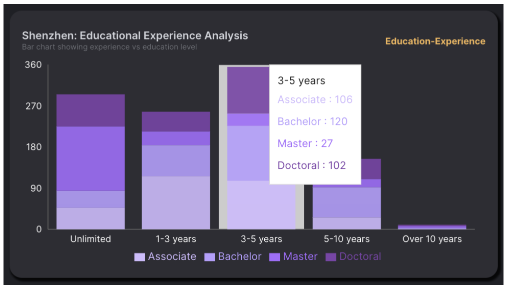
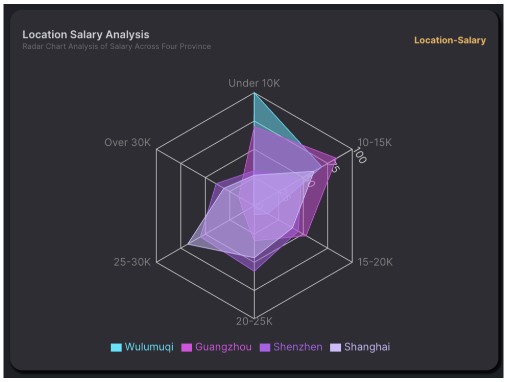


Figure 7—Bar Chart

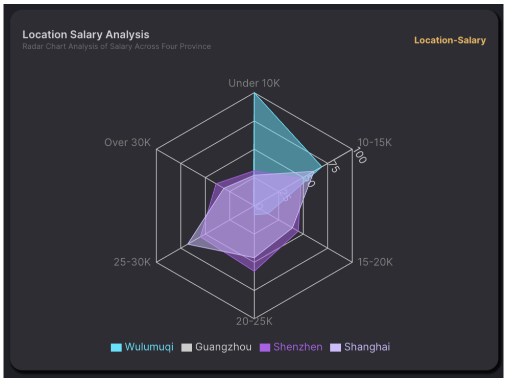


Location Salary Analysis

Radar Chart Analysis of Salary Across Four Province

Location-Salary

Figure 8—Radar Chart

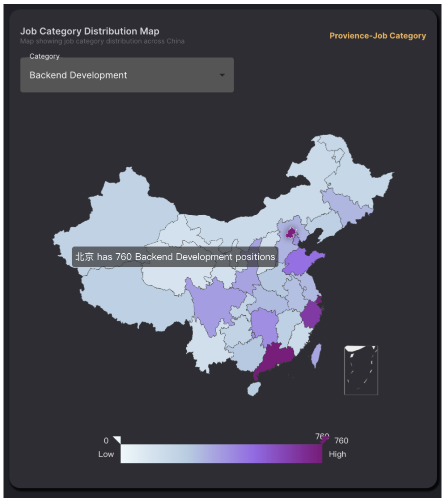


Location Salary Analysis

Radar Chart Analysis of Salary Across Four Province

Location-Salary

Figure 9—Radar Chart: hide Guangzhou



Job Category Distribution Map

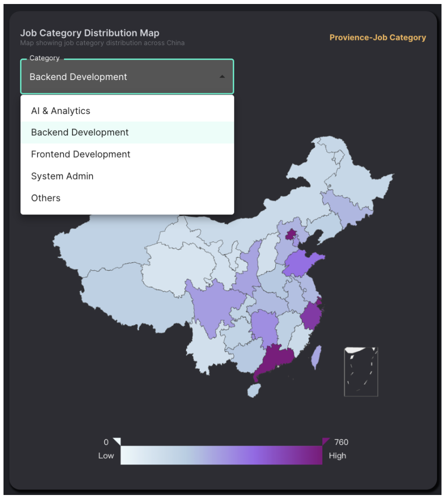
Map showing job category distribution across China

Proviencе-Job Category

Category: Backend Development

北京 has 760 Backend Development positions

Figure 10—China Map



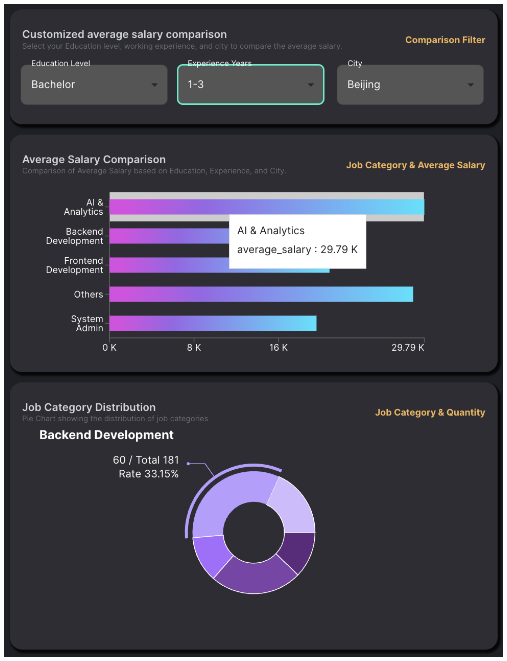
Job Category Distribution Map

Map showing job category distribution across China

Provience-Job Category

Category: Backend Development

Figure 11—China Map Select



Customized average salary comparison

Select your Education level, working experience, and city to compare the average salary.

Comparison Filter

Education Level: Bachelor

Experience Years: 1-3

City: Beijing

Average Salary Comparison

Comparison of Average Salary based on Education, Experience, and City.

Job Category & Average Salary

AI & Analytics

average_salary : 29.79 K

Job Category Distribution

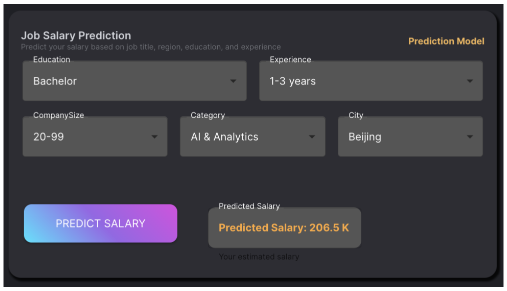
Pie Chart showing the distribution of job categories

Job Category & Quantity

Backend Development

60 / Total 181
Rate 33.15%

Figure 12—User Customized Filter



Job Salary Prediction

Predict your salary based on job title, region, education, and experience

Prediction Model

Education: Bachelor

Experience: 1-3 years

CompanySize: 20-99

Category: AI & Analytics

City: Beijing

Predicted Salary: Predicted Salary: 206.5 K

Your estimated salary

Figure 13—Prediction