PREDICTING SALARIES FOR DATA-RELATED JOBS

BEHRAD KHAMENHMOHAMMADI
(STUDENT ID: 1251322633)
MSC IN IT FOR BUSINESS DATA ANALYTICS

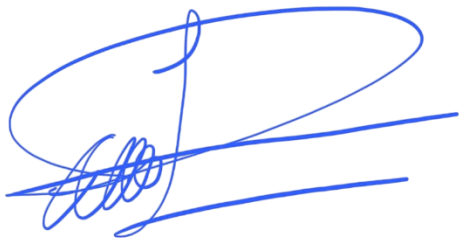
DISSERTATION SUBMITTED TO INTERNATIONAL BUSINESS SCHOOL FOR THE PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN IT FOR BUSINESS DATA ANALYTICS

<<December 2025>>

DECLARATION

This dissertation is a product of my own work and is the result of nothing done in collaboration.

I consent to International Business School's free use, including online reproduction, including electronically, and including adaptation for teaching and education activities of any whole or part item of this dissertation.



Behrad Khamenhmohammadi

Word length: 10,001 words

EXECUTIVE SUMMARY

This project examined how salaries in data-related jobs change across different countries, levels of experience, and work conditions. The main focus was to understand the factors behind salary differences and to build a machine-learning model that can estimate salary ranges. The study used a large international dataset covering the years 2020 to 2025, and several new features were created to make the dataset more suitable for analysis, such as seniority level, remote-work category, and indicators for post-COVID and post-AI periods.

In my project several models were tested, and the Random Forest model showed the most stable performance, even though the overall accuracy remained modest. The model's behaviour was interpreted using SHAP, which helped identify the features that influence predictions the most. Seniority level was the strongest driver, followed by the employee's country of residence. As in previous researches which found experience and geography are important factors, in my research support this idea as well in salary variation.

The project also analysed salary trends over time. The results showed a clear increase from 2022 onward, with 2023–2025 showing the highest levels. These patterns match global changes in the job market, including the long-term impact of COVID-19, the rise of remote work, and the growing influence of AI tools such as ChatGPT.

To check how well the model generalises, it was tested on an external dataset. The model showed similar patterns but with larger errors because the external dataset had different structures and salary distributions. This confirmed both the strengths and limits of the model.

Overall, the project provides practical insights for HR teams, employers, and job seekers. While salary prediction cannot be fully precise, the results help identify the main drivers of salary differences and offer a clearer picture of how the data job market has changed from 2020 to 2025.

Table of Contents

LIST OF TABLES

table page

LIST OF FIGURES

figure **page**

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND

The job market for data-related work is changing very fast. Many companies need data analysts, data scientists, and AI engineers to help them work with information and make better decisions. Because these jobs are new and very different from traditional office jobs, people often do not know what a “fair salary” should be. Salaries can also be very different between countries, companies, and job titles. For example, a data scientist in one country may earn much more than a data analyst in another place, even if they have similar skills.

In the last few years, many global events also changed the job market. The pandemic has made remote working more common around the world, and more companies have adjusted their salary levels in this period of time. New technologies, such as the rise of artificial intelligence, have also changed skill requirements, which has had a significant impact on salary expectations. With these changes, it is important to understand the factors that affect salary levels in all occupations, including data-related occupations. A clear view of these factors can help job seekers make better decisions and help employers create fair and competitive salary plans.

1.2 PROBLEM STATEMENT

Salary levels in data-related jobs are very different across countries, companies, and job roles. Some studies show that factors such as company size, company location, job title and experience can create strong differences in salaries. Because of this, job seekers often do not know what a fair salary is, and employers also find it difficult to set the right pay.

Global events also created more problems for the labour market. The COVID-19 pandemic changed skill needs and salary patterns in many jobs. Semenenko et al. (2024) also show that the Russia–Ukraine war caused serious economic instability in Russia, which can influence growth, living standards and job opportunities. New AI tools, such as ChatGPT, also changed job structures and created uncertainty in many occupations.

Because of all these factors, it is difficult to understand which variables have the biggest effect on salaries in data-related jobs. There is a need for a clear model that shows the important features and explains how external events may change salary patterns.

1.3 RESEARCH AIM AND OBJECTIVES

RESEARCH AIM

The aim of this project is to build and interpret a machine learning model that predicts salaries for data-related jobs, identifies the most important factors that influence salary levels, and examines how external events between 2020 and 2025 — such as COVID-19, the Russia–Ukraine

war, and the rise of AI tools like ChatGPT — may have changed salary patterns in the data job market.

RESEARCH OBJECTIVES

To achieve to this point, the project has the following objectives:

1. **To explore key factors that influence salary levels** in data-related jobs.
2. **To clean and analyse the main dataset** using exploratory data analysis (EDA) to understand general salary patterns by year, experience level, job title, location, company size, and remote ratio.
3. **To build and compare machine learning models** (Linear Regression, Decision Tree, Random Forest) for predicting salaries and to tune their hyperparameters for better performance.
4. **To interpret the models** by identifying important features using feature importance methods and SHAP values.
5. **To analyse salary changes during major external events**, including COVID-19, the Russia–Ukraine war, and the rise of AI tools (such as ChatGPT), and to compare normal years with shock years.
6. **To validate the final model** by testing it on an external dataset (such as the StackOverflow Developer Survey) and comparing the prediction results.

1.4 RESEARCH QUESTIONS

Based on the project's aim, the following research questions guide the study:

1. What are the main factors that influence salary levels in data-related jobs?
2. How do salary patterns change across different years, job titles, experience levels, company locations, and company sizes?
3. Which machine learning model can best predict salaries for data-related jobs?
4. Which features are most important in the prediction model, based on feature importance and SHAP values?
5. How did major external events between 2020 and 2025 — such as COVID-19, the Russia–Ukraine war, and the rise of AI tools like ChatGPT — affect salary trends in the data job market?
6. How well does the final model perform when tested on an external dataset?

1.5 SCOPE AND SIGNIFICANCE

SCOPE

This project focuses on analysing and predicting salaries for data-related jobs using a structured dataset from 2020 to 2025. The study includes variables such as job title, experience level, company location, company size, employment type, remote ratio, and salary. The project does not include web scraping, software deployment, deep learning models, or real-time applications. Instead, it focuses on traditional machine learning models, model interpretation, and analysis of salary changes during major external events, such as COVID-19, the Russia–Ukraine war, and the rise of AI tools like ChatGPT. The scope also includes testing the final model on an external dataset to check its reliability.

SIGNIFICANCE

This research is significant because salary levels in the data job market are changing quickly and are influenced by many factors. Job seekers often do not know what a fair salary is, and employers may find it difficult to set competitive pay. By building a clear and interpretable machine learning model, this study helps both groups understand how different factors affect salaries. The project also shows how external events can change salary trends, which is important for companies and policy makers. The results can support better decision-making, improve transparency, and help people plan their careers in the data industry.

1.6 OVERVIEW OF ANALYTICAL APPROACH

This project follows a step-by-step analytical process. First, the dataset is cleaned and explored to understand salary patterns. Then, some machine learning models such as Linear Regression, Decision Tree and Random Forest are trained and tuned. The models are interpreted using feature importance and SHAP, and finally validated with an external dataset to check reliability.

1.7 STRUCTURE OF THE DISSERTATION

This dissertation is organised into nine chapters. Chapter 1 introduces the topic, the problem, and the goals of the study. Chapter 2 reviews the academic literature and explains the main factors that influence salaries. Chapter 3 connects business needs with analytics. Chapter 4 describes the dataset and presents the exploratory analysis. In chapter 5 we are going to explain the methodology and the machine learning models. Chapter 6 shows the implementation in Python. Chapter 7 validates the model with an external dataset. Chapter 8 discusses the results. Finally, Chapter 9 presents the conclusion and future work.

CHAPTER 2
LITERATURE REVIEW

2.1 SALARY DETERMINANTS

Salary levels are influenced by several groups of factors. We can collect these factors such as Organizational factors, Job-related factors, Individual factors, Economic and market factors, and Social and cultural factors, which are explained as below:

Organizational factors

Organizational factors include company size, location, industry and pay structure. Many salary reports say that bigger companies or companies in profitable industries usually offer higher pay. Recent research also supports this idea. For example, Khosa et al. (2024) use the Global Salary 2024 dataset and find that company location and size have a strong effect on entry-level salaries in global labour markets.

Job-related factors

These factors include job complexity, responsibility, and skill requirements. In the factor list, jobs with higher responsibility or expert skills usually receive higher pay. This point is also seen in Lu et al. (2025), who found that technical skill requirements in big-data roles strongly influence salary levels in job advertisements.

Individual factors

Many studies show that individual factors such as education, work experience, technical skills, and gender can influence wages. These personal characteristics can affect how much people earn in both lower-paid and higher-paid positions.

In the medical sector, Ramkumar et al. (2024) found a clear gender pay gap. Even after controlling for case volume, practice type, and years in practice, women surgeons still earned about 14% less than men.

Economic and market factors

These include labour supply, economic stability, inflation, and labour laws. For example, Goerke and Pannenberg (2025) show that minimum-wage compliance depends on economic and legal conditions inside firms.

Social and cultural factors

These include fairness, social norms, and gender dynamics. De la Torre-Ruiz et al. (2024) show that how pay information is communicated affects employees' pay satisfaction and their feeling of support from the organisation.

2.2 LABOR MARKET TRENDS (2020–2025)

COVID-19 Impact (2020–2022)

COVID-19 changed labour markets in many countries after it is happened. In a study from Mexico, it was written that COVID-19 reduced labour market participation for women with children and increased job instability (Juarez & Villaseñor, 2024). This shows that a large crisis can affect job access, working hours, and also salary patterns.
During this pandemic period, digital workings is became more common, and many companies moved to remote work. As a result, skills related to data and technology became more important, and this increased demand for workers in these areas. These changes influenced salaries in data-related jobs as well.

Russia–Ukraine War (2022–2025)

The Russia–Ukraine war brought a lot of economic uncertainty to the region, and this also affected how local labour markets worked (Semenenko et al., 2024). When the economy becomes unstable, salaries in some sectors may change as well, mainly in countries that are close to the conflict or have strong trade links with it. International tech companies can also feel these changes because exchange rates, inflation, and investment plans become less predictable. Since the dataset in this project includes both year and country, it gives a chance to see how salaries moved during this time and to connect those changes with these outside events.

Remote Work Changes (2020–2025)

Digital skills became essential after COVID-19, especially in remote-friendly sectors. This approach means that workers with strong soft and hard skills together could find more job opportunities with more salaries. Because the main dataset includes **remote_ratio** and **company_size**, these factors can be analysed directly in this project.

US Elections and Political Shifts (2020 & 2024)

The technology sector in the United States is important for global labour markets, because many AI and data-driven companies are based there. Changes in US policy can influence how companies adopt AI, which jobs grow, and which skills become more or less important. Political changes can affect hiring, investment levels, and therefore salary patterns.
When comparing different years in the dataset (for example 2020–2025), some salary changes may be linked to these political events.

2.3 IMPACT OF AI & CHATGPT ON DATA SCIENCE SALARIES

The fast growth of Artificial Intelligence (AI) has changed many job markets, including data-related roles. AI brings both new opportunities and new risks for salaries in technical jobs. AI automation reduces some routine work, but at the same time increases the value of advanced technical skills. This means that some job roles may lose salary power, while others may become more valuable.

A noticeable change appeared after tools like ChatGPT entered the workplace. In Zarifhonarvar's study (2024), he explains that these systems may fully influence a large group of jobs and partly affect many others. The main reason is that tasks based on writing, reading or basic analysis can now be done much faster than before. Because of this shift, some mid-level positions may face slower salary growth, while roles that rely on judgement, experience or decision-making usually keep a stronger salary level.

Student perception studies also support this view. Gomes and Brito (2025) show that students in technical fields see AI as helpful for learning and productivity, but they also worry that some traditional roles may disappear. This supports the idea that salary trends may shift depending on how companies use AI tools.

AI exposure changes wage patterns differently across industries. High-skill technical jobs may see wage increases, while jobs with repetitive tasks may face slower growth. Together, these studies show that AI and ChatGPT do not have a single effect. Instead, they create a mixed impact where skill level, job role, and company strategy decide how salaries change in data-related jobs.

2.4 ML APPROACHES IN HR ANALYTICS & SALARY PREDICTION

Machine learning is widely used in HR analytics because it helps organisations understand patterns in salaries and make better decisions. Several studies show that ML models can analyse many job-related and personal factors at the same time and estimate pay levels with good accuracy.

Ayua et al. (2023) used *polynomial regression* to predict salaries of non-academic staff. Their work showed that non-linear models can capture complex relationships between experience, job level and salary (Ayua, Malgwi and Afrifa, 2023). This supports the idea that salary patterns are not always linear and may depend on combinations of factors.

Other researchers also used multiple linear regression to study salary patterns in technology jobs. In the study by Maidin and Ayyasy (2025), the model used job location, experience level and technical skill information to predict salaries for AI professionals. Their findings show that location and experience have a strong effect on salary, while technical skills only add a small improvement. The model still explained a large part of the salary differences, which supports the use of linear models for this type of analysis.

Machine learning is also used to explore general salary trends for data jobs, and it applied *linear regression* to a data-science salary dataset and highlighted how variables such as job title and experience level affect salary. Some studies reported that simple ML models can still provide useful predictions when the dataset is large and well-structured (Deb, 2025).

More advanced work shown the importance of optimisation and hyperparameter tuning. These workings are studied different regression models in scikit-learn and showed that tuning hyperparameters significantly improves model performance. Results suggest that optimisation steps, even for basic algorithms, can reduce errors and produce more stable predictions (Salama, 2024).
Together, these studies show that ML techniques such as linear regression, polynomial regression, and tuned regression models are widely used in HR analytics and salary prediction.

The literature suggests that models become more accurate when:

- the choice of features is meaningful,
- hyperparameters are tuned,
- and model complexity matches the structure of the data.

These insights support the design of this project, where multiple ML models will be tested, feature importance will be analysed, and hyperparameters will be optimised to improve prediction quality.

2.5 SUMMARY OF GAPS IN THE LITERATURE

Many studies talk about salary differences and the things that can change them. They explain how experience, job title, company size or location can affect pay. But most of these studies look at one country or one type of job. So we still do not have a clear and wide view of salary patterns in data-related jobs in different parts of the world.
There are also studies about COVID-19 and the Russia–Ukraine war, but they mainly focus on general labour market changes. They do not show clearly how these events influence salary prediction for data jobs. Research about AI and ChatGPT also exists, but the real effect on salaries in data roles is still not fully explained.
In machine learning studies, many papers use simple models. Only a few combine several models, feature engineering, model interpretation, and testing with an external dataset. Because of these missing parts, there is still space for a study that uses a stronger ML approach and checks the results more carefully.

CHAPTER 3
BUSINESS ANALYTICS INTEGRATION

3.1 STAKEHOLDERS

In salary prediction for data-related jobs, several groups are involved, and each group needs different types of information. The first group is job seekers. They want to know if the salary they receive is fair compared to others with the same skills or experience. Many people in the labour market face uncertainty because salary levels are very different across companies and countries, and this makes decision-making difficult. Clear predictions can help them understand their value better.

The second group is HR teams. HR departments often need to plan budgets, decide pay ranges, and reduce unfair differences in salaries. Salary differences can come from many factors such as experience, job role, company size, and work conditions. HR teams can use prediction results to check if the salary they offer is competitive or if changes are needed.

The last group is companies and managers. Companies want to attract skilled employees, especially in data jobs where the demand is high. Forecasting salary levels can help employers to design better hiring strategies and understand market changes, especially during events like COVID-19 or economic surprises. In general, all these groups can benefit from a model that explains how salary patterns change and which factors matter most.

3.2 BUSINESS QUESTIONS & DECISION CONTEXT

In companies that work with data jobs, making decisions about salaries is not always easy. The job market changes very fast, and the required skills also change from year to year. Because of this, managers and HR teams often face a few important questions.

One question is about the factors that create salary differences. For example, they want to know if the job title is more important, or the experience level. They also wonder if the size of the company or the company location has a stronger effect. Companies need this information to offer fair and correct salaries.

Another question is about the influence of external events on the job market. During the COVID-19 period, many jobs changed. Later, with new AI tools such as ChatGPT, some skills became more valuable and some tasks became easier. Companies want to understand if these changes increased or decreased salaries in data-related roles.

Organizations also want to know which skills help employees earn higher salaries, and if remote work has changed salary levels. This is important for HR planning because many companies are still deciding between office work, hybrid work, and remote work.

Finding answers to these questions helps companies make better decisions about hiring, salary budgets, and employee retention. This project uses data analysis and machine learning to give

clear information about these topics. The goal is to show where salaries are aligned with the market and where companies may need to review their salary strategy.

3.3 HOW SALARY PREDICTION SUPPORTS HR STRATEGY

Salary prediction models can help HR teams make decisions in a more clear and structured way. In many companies, HR managers must decide how much to offer for new positions, how to adjust salaries for current employees, and how to plan salary budgets for the next year. These decisions are difficult when there is not enough information about the job market.

A prediction model gives HR a simple tool to understand what a fair salary looks like for a specific role. When HR enters job details such as job title, experience level, company size, or location, the model can estimate a realistic salary range. This helps companies avoid offering salaries that are too low or too high.

Another benefit is fairness. If similar jobs inside the company receive very different salaries, employees may feel that the system is not fair. A prediction model can show HR which jobs are outside the normal market range. This helps HR adjust salaries in a more transparent and objective way.

Salary prediction also supports planning. HR teams need to prepare budgets for new projects, new teams, or expansions into other countries. With a prediction model, they can see how salaries change in different regions or job roles. This makes long-term planning easier and reduces financial risk.

The model can also help HR understand how external events may affect salaries. For example, during COVID-19 or after the appearance of new AI tools, some roles became more expensive because the demand increased. By looking at salary trends from 2020 to 2025, HR can prepare better for future changes in the labor market.

Overall, salary prediction supports HR strategy by giving clear, data-based information. Instead of relying only on intuition or past salary tables, companies can use real data to make decisions that are fair, consistent, and aligned with market conditions.

3.4 SUCCESS CRITERIA & VALUE CREATION

The success of this project is not only about building an accurate model. It also depends on whether the results can be useful for the people who will use them. For this reason, several criteria are considered to measure the real value of the model.

The first criterion is prediction accuracy. The model should estimate salary levels with a reasonable level of error. Metrics such as MAE and R^2 help to check how close the predictions are to real values. After accuracy, clarity of the results is important. It is not enough to give a number; users should also see which factors—such as experience, job title, company size, or location—had the strongest influence on the prediction. Tools like feature importance and SHAP make this easier to understand.

Another important criterion is practical usefulness. If HR teams can use the model to check whether their salary offers match the market, or if job seekers can understand what affects their expected salary, then the model has real value.

The model should also prove that it works not only on the original dataset. Testing it on an external dataset shows whether it performs well in real-world situations. This is especially important when the job market changes because of major events such as COVID-19, the Russia–Ukraine war, or the rise of AI tools.
At the end, when the model help to make better decisions about salaries, recruitment, and workforce planning, and if it helps individuals understand the factors that shape their earnings, we can say the model is successful.

3.5 LIMITATIONS OF REAL-WORLD SALARY DATA

Real salary data always comes with some gaps. Many datasets only show the main salary and leave out things like bonuses or other benefits that people get during the year. Because of this, the numbers we see are not always the full picture of a person's income.
Job titles are another problem. Companies often use their own naming style, so the same title might mean different tasks in different places. Sometimes two jobs look different on paper but are almost the same in real work. This makes comparing roles across companies a bit confusing.
There are also issues inside the data itself. Some records have missing parts, and some information is entered by people manually. This can lead to small errors, or sometimes the data does not match perfectly with what happens in real life.
Salary levels also change because of things that are not written in the dataset. Events like inflation, political decisions, or sudden changes in the market can affect salaries, but the dataset cannot show these reasons. We only see the numbers going up or down without knowing the background.
Besides these, some personal factors—such as negotiation skills or the rules inside a company—cannot be measured in a simple file. These things matter in real salary decisions but stay invisible when we look only at the data.

CHAPTER 4
DATA & EXPLORATORY DATA ANALYSIS

4.1 DATASET DESCRIPTION

The dataset used in this project is named main_salary_dataset.csv. It includes 93,597 job records and 11 columns, and each row shows one position from the data-related field. The dataset covers the years 2020 to 2025, stored in the work_year column.
The main variable for prediction is salary_in_usd, which gives the annual income in US dollars. The original salary and currency also appear in the column's salary and salary_currency, so the data allows both local and standardised comparisons.
Several columns describe the job itself. The experience_level column has four groups, from entry level to executive, and employment_type shows whether the job is full-time, part-time, contract, or freelance. There are 317 job titles, and "Data Scientist" appears most often.
The dataset also includes basic location and company details. The worker's country is stored in employee_residence, while company_location shows the firm's country. The remote_ratio column explains how much of the job is done remotely (0, 50, or 100). Company size is grouped into three categories. All columns are complete, and there are no missing values.

4.2 DATA CLEANING & PREPROCESSING

Before starting the analysis, the dataset was checked to confirm that it was ready for modelling. The file contains 93,597 rows and 11 columns, and all of them were loaded without any issues. A first check showed that there are 46,960 duplicated rows, which is a large amount. Keeping these rows would distort the results, so they need to be removed before building the models.
The dataset also was examined for missing values, and the results showed that no column contains missing data. This makes the preprocessing simpler because no imputation is required. The column types were also reviewed: four columns are numerical, and the others are categorical, such as job title, experience level, and company location. These categorical fields will later be encoded so the machine-learning models can process them.
A short look at the salaries showed a wide range, from 15,000 to 800,000 USD, which suggests the presence of outliers. These values will be examined again during the visual exploration stage. Overall, the dataset is complete, but it needs duplicate removal and preparation of the categorical fields.

4.3 Summary Statistics

THE salary variable in the dataset shows a wide spread. The average salary in USD is about 157,548, with a median of 146,232. Most salaries are between roughly 106,250 and 198,000 USD, but a few records go up to 800,000 USD.

There are 317 different job titles, which shows that the data field is very diverse. The most frequent roles are Data Scientist, Data Engineer and Software Engineer. Experience level also affected the range of salary. Entry-level employees earn the lowest average salaries, and executive-level employees earn the highest. These results show brilliantly that both job role and experience have a strong impact on salary in this dataset.

4.4 VISUAL EDA

In this part of the analysis, visual charts are used to better understand the salary data and the main patterns inside the dataset. By showing the information in graphs, it becomes easier to see changes across years, differences between experience levels, and variations across locations or remote work levels.

These visuals help to identify trends, unusual values, or patterns that may influence the final machine learning model. Each figure is followed by a short explanation that describes what can be observed from the chart.

Figure 4.1 shows how many job records appear in the dataset for each year from 2020 to 2025. The counts increase sharply after 2022, with the largest share coming from 2024. This pattern reflects the strong growth in data-related job reporting during recent years.

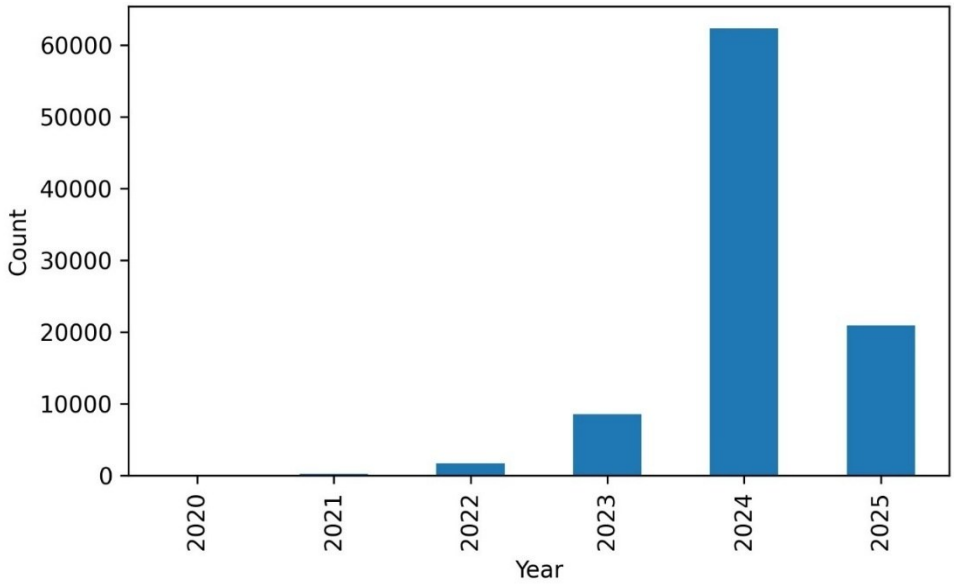


FIGURE 4.1 NUMBER OF RECORDS BY WORK YEAR

Figure 4.2 shows how the average salary changed between 2020 and 2025. The values increase steadily from around 100,000 USD in 2020 to a peak in 2024. The slight drop in 2025 suggests a minor correction after several years of continuous growth in data-related job salaries.

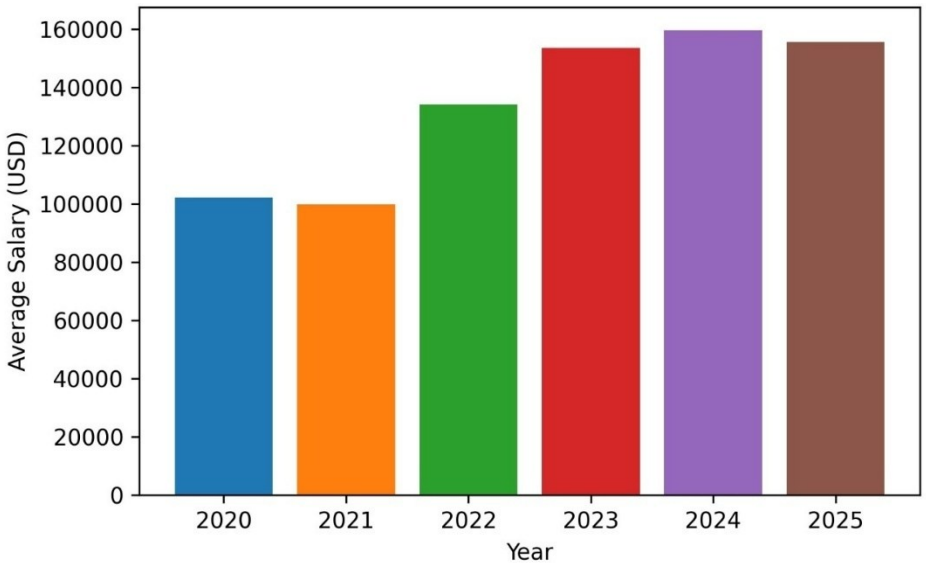


FIGURE 4.2 AVERAGE SALARY IN USD BY WORK YEAR

Figure 4.3 shows how the remote work ratio is divided in the dataset. Most records belong to jobs that are completely on-site, and only a smaller part of the data relates to fully remote roles. The 50% group is very small and almost not visible in the chart.

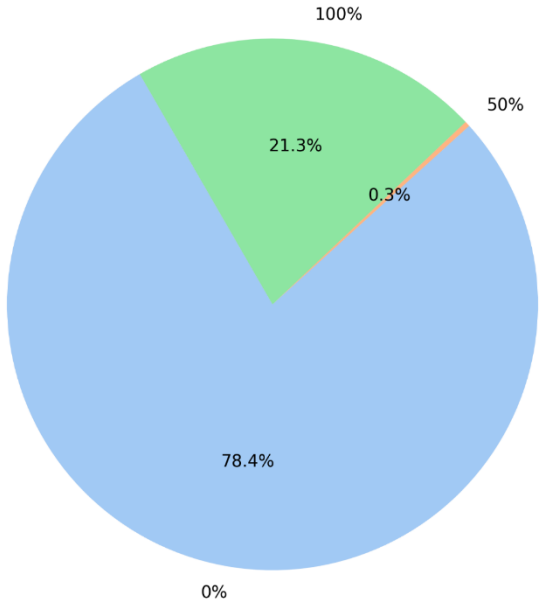


FIGURE 4.3 DISTRIBUTION OF REMOTE RATIO

This means that the dataset mainly includes two types of jobs: either people work at the office or they work fully from home. There are not many hybrid jobs here. Because of this, the remote_ratio field is not very balanced and may not have a strong effect when we later train the prediction models.

The results in Figure 4.4 show that the dataset is heavily dominated by workers living in the United States. More than 84,000 records come from the US, while the second and third positions (Canada and the United Kingdom) are far behind, with only a few thousand entries. The remaining countries appear with very small counts.

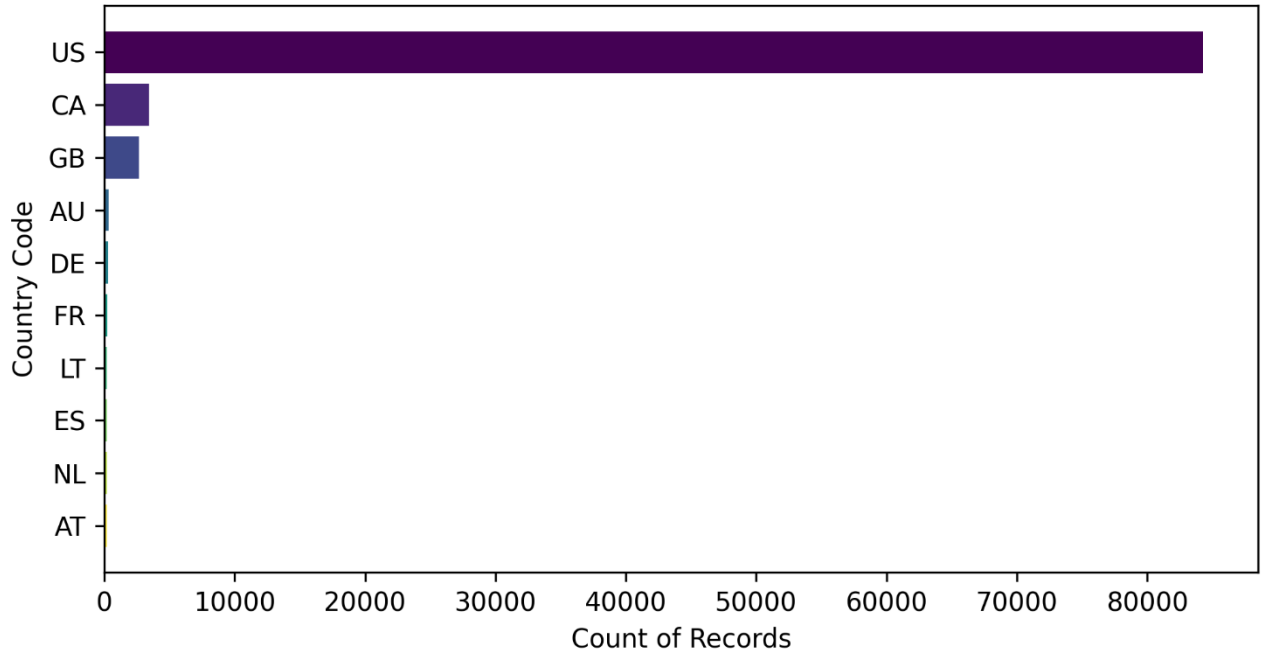


FIGURE 4.4 TOP 10 EMPLOYEE RESIDENCE COUNTRIES

This uneven distribution suggests that the dataset mainly reflects the North American technology labour market. Because of this concentration, many patterns observed in later analysis—such as salary levels or remote-work behaviour—will naturally be influenced by the US market. This point needs to be kept in mind when interpreting the findings in the next chapters.

4.5 MAPPING DATASET FACTORS TO LITERATURE FACTORS

When comparing the dataset with the main groups of salary factors mentioned in the literature, some parts match well while others are missing. The job-related factors are the clearest ones. We have job title, experience level, and employment type, and these already show different levels of responsibility or skill in the workforce. From the organisational side, company size is included, and it gives some idea about how salaries may differ between small, medium, and large firms. The location of the company can also be linked to living costs or general market conditions, even though it is not explained in detail in the dataset.

However, several important factors are not present. The data does not include education level, certificates, or personal skills, so individual differences cannot be studied properly. Wider economic and social influences—such as inflation, labour rules, or fairness issues—are also not

part of the dataset. Because of these missing elements, the data cannot fully reflect all salary drivers, but the variables that are available still help us understand the main structural patterns in data-related jobs.

4.6 DATA LIMITATIONS & BIAS

Our main dataset still has several limits that can influence the analysis. One of the biggest is the high number of duplicated rows. When many records appear more than once, some job titles or salary levels become over-represented, and this can push the results in a certain direction. Another limit is the wide spread of job titles. Some titles have thousands of records but others only a few, so comparisons across all roles are not always balanced.

The dataset also misses many personal factors that usually affect salaries, such as education level, certificates, or specific technical skills. Because of this, the model can only work with general patterns and not with deeper individual differences. The same happens with company information: we know the company size, but we do not know the industry, internal pay policy, or financial strength, which are usually important in salary studies.

There is also a geographical bias. Most records come from the United States, while many other countries have very small samples. This means the results of the model will naturally be closer to US salary patterns and may not represent other regions well. Finally, some extreme salary values exist in the data, and even though they are real entries, they may influence model behaviour if not handled carefully.

Overall, the dataset is useful for building salary prediction models, but the missing variables and imbalanced distributions can limit how widely the results can be interpreted.

CHAPTER 5
METHODOLOGY

5.1 Feature Engineering

At this stage, several new features were created to help the model capture salary patterns more accurately. First, duplicate rows were removed so that the modelling process relies only on unique observations. After this step, the number of records decreased from 93,597 to 46,637.
To represent experience differences in a clearer way, a numerical feature called seniority_level was added. This feature converts the four experience categories (EN, MI, SE, EX) into ordered values, which makes it easier for the model to compare different levels.
Remote-work information was also reorganised into a more descriptive feature. The new variable remote_category groups jobs into three types: on-site, hybrid, and fully remote. This helps in analysing how working location relates to salary differences.
Another feature, same_country, indicates whether the employee's residence and the company's location are in the same country. This is useful for understanding salary differences between local and cross-border employment. Two time-based features were added as well.
Post-COVID marks the years after the COVID-19 period (2021 onwards), and post-ai marks the years after the release of ChatGPT (2023 onwards). These variables help the model capture salary shifts connected to major global events.
Overall, these engineered features will support the model in identifying clearer patterns during the prediction tasks.

5.2 ENCODING, SCALING & TRAIN/TEST SPLIT

After creating the new features in the previous step, the next task was to prepare the data for the machine-learning models. For this part, only the cleaned dataset (df_model) was used, because the original file contained many duplicate rows. The modelling dataset includes ten features, such as the work year, seniority level, remote work ratio, country information, company size, and the indicators we added for post-COVID and post-AI years.
Since the dataset contains both numeric and categorical variables, they could not be used directly in a single model. The numeric features were standardised so that they would be on a similar scale, which helps algorithms that are sensitive to large value differences. The categorical features are transforming with one-hot encoding so that each category becomes a binary column.
After the preprocessing steps were defined, the data was divided into training (80%) and testing (20%) parts. The final split resulted in 37,309 rows for training and 9,328 rows for testing, which provides a stable base for the upcoming model comparison.

5.3 MACHINE LEARNING MODELS

Three classical regression models were trained to predict salaries in USD. All these models were trained on the same processed data, using scaled numeric features and one-hot encoded categorical variables. The first model is a Linear Regression, which served as a simple baseline. Its performance on the test set showed a MAE of about 51,132 USD and an R^2 value of 0.19, meaning that it explained only a small portion of the variation in salaries.
The second model was a Decision Tree Regressor. Decision trees are able to capture non-linear relationships, so they usually perform slightly better than linear models. In this case, the tree produced a MAE of around 50,867 USD with an R^2 of 0.19, which is only a minor improvement over the linear model. The limited gain suggests that the dataset has a high level of variation that a single tree cannot fully capture.
The third model was a Random Forest Regressor, which averages many trees to reduce overfitting. Random Forest model achieved the best results between these three models, with a MAE of 50,764 USD and an R^2 of 0.19. Although the improvement is small, the model is more stable and provides a better base for the next steps, such as hyperparameter tuning and feature importance analysis.

5.4 HYPERPARAMETER TUNING

To upgrade the performance of the Random Forest model, several different parameter combinations were tested. In this step, the number of trees, the depth of the trees, and the conditions for splitting each node were adjusted to see which setup worked best. The tuning was done using cross-validation so that the results would not depend on only one random train–test split.
After running all parameter combinations, the version with 200 trees, no fixed depth limit, and the standard values for minimum samples per split and leaf performed the best. The mean absolute error in this case was around 51,000 USD. Although the improvement over the initial model was not very large, the tuned model showed more stable behaviour and its results fluctuated less. Because of this consistency, the tuned Random Forest model was chosen for the next steps of the project.

5.5 MODEL INTERPRETATION TOOLS

To understand how the model makes its predictions, two interpretation tools were applied: the feature importance values produced by the Random Forest model and SHAP values, which provide a more detailed explanation of feature influence.
Table I. shows the top features ranked by their mean absolute SHAP values. The results make it clear that seniority_level is the strongest predictor in the model. This means that, in this dataset, the employee's level in the job hierarchy has the largest effect on salary changes. The second most important factor is employee_residence_US, which suggests that jobs linked to the United States show noticeably higher salary levels.

TABLE I. SHAP- FEATURE IMPORTANCE

	feature	mean_abs_shap
1	seniority_level	19943.732066
95	employee_residence_US	14908.607816
113	company_location_CA	1688.374536
0	work_year	1663.684386
7	remote_category_on-site	1533.404872
2	remote_ratio	1456.521744
21	employee_residence_CA	1293.364009
40	employee_residence_GB	1033.940052
183	company_location_US	818.108370
8	remote_category_remote	587.070268

A smaller group of features also has a visible but weaker influence. These include company_location_CA, work_year, and the remote_category_on-site variable. Their presence in the top positions indicates that geography and work arrangement still shape salary outcomes, although they do so to a smaller extent compared with job seniority.

The SHAP summary plot (Figure 5.1) visually supports this pattern. In the plot chart, the longest bars belong to seniority_level and employee_residence_US, meaning these variables shift the model's predictions more than others. The remaining features appear with much shorter bars and they show that their effect is low on the salary rate.

Based on both Table I. and Figure 5.1, show that the model's predictions mainly follow structural job factors (such as seniority) and country-level differences (especially the U.S. market). Other features such as remote work ratio or company size still matter but contribute less to the final salary estimates.

5.6 ETHICAL NOTES

When working with salary data and predictive models, several ethical points must be considered. First of all, salary predictions should not be used to justify unfair pay decisions. The goal of this project is to support transparency, not to replace human judgement. Second, the dataset contains differences between countries, job levels, and remote-work categories. These patterns may reflect existing inequalities in the labour market, and the model may reproduce them if they are not interpreted carefully. Third, because some job groups appear much more often than others, the model naturally becomes more accurate for larger groups. This limitation must be communicated clearly to anyone who uses the results. Finally, no personal identifiers were included in the dataset, which reduces privacy risks, but any real-world application should follow standard data-protection rules.

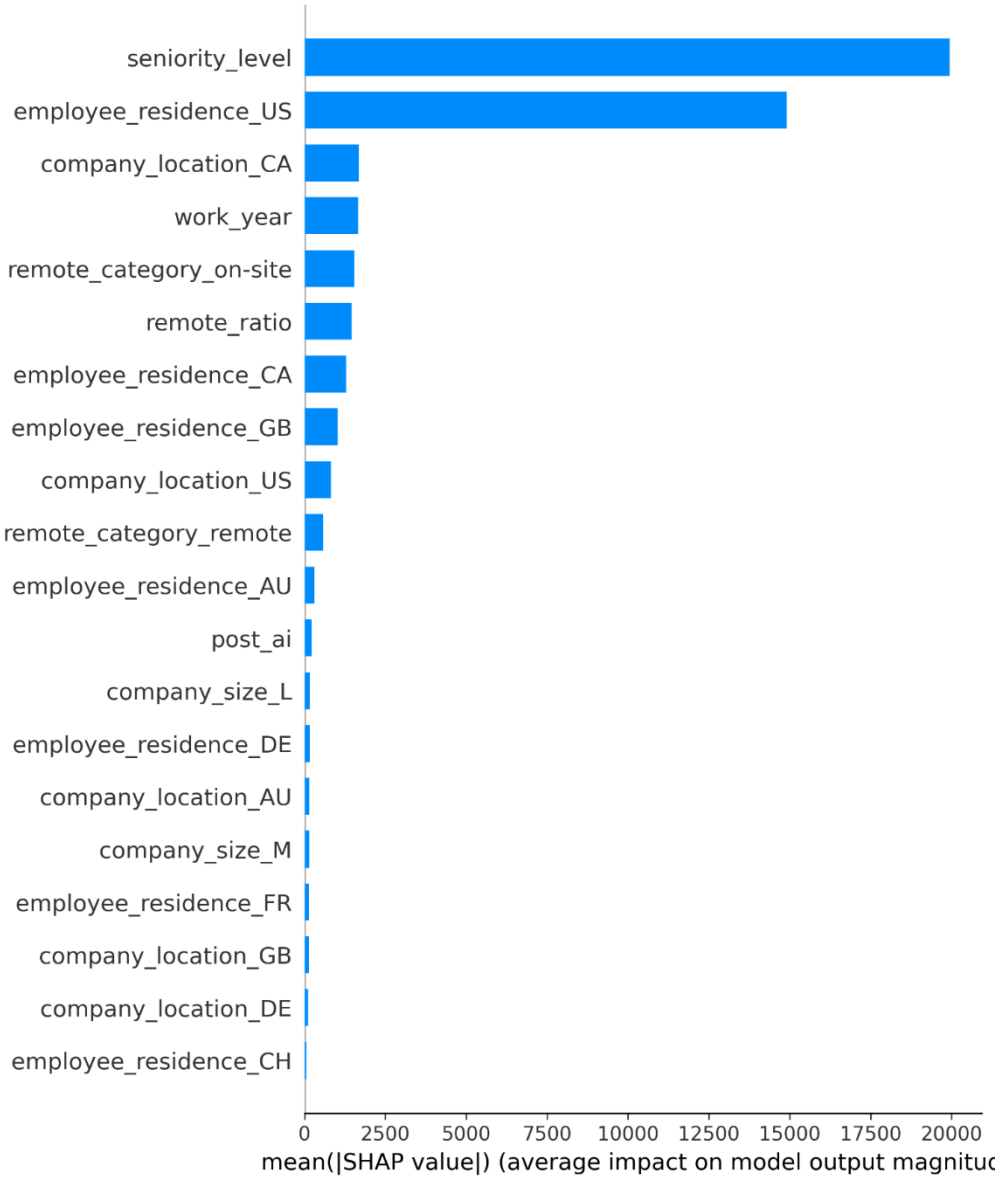


FIGURE 5.1 SHAP - FEATURE_IMPORTANCE

CHAPTER 6
IMPLEMENTATION IN PYTHON

6.1 ENVIRONMENT (COLAB) & REPRODUCIBILITY

All parts of project was implemented in Google Colab, which provides a cloud-based Python environment with all required libraries already available. Using Colab helped keep the work reproducible, because the code runs the same way on any device without needing a local installation. A fixed Python version and a stable set of packages were used, and these packages were also listed inside the project's requirements.txt file. This makes it possible for anyone to recreate the same environment if they want to run the notebook outside Colab.

All steps of the analysis—loading the dataset, cleaning, feature engineering, training the models, and generating the figures—were completed inside a single notebook called salary_analysis_main.ipynb. Running the notebook from the first cell to the last one produces the same outputs every time, including the saved charts and the trained model results.

To support full reproducibility, all figures and important outputs (such as SHAP results) were exported and stored in the Figures folder in the GitHub repository. This means that even if the notebook is not executed, readers can still see all generated results. Using a public GitHub repository also ensures transparency: both the raw data and the processed results are available, and all steps of the analysis can be reviewed or repeated by other researchers.

6.2 PROJECT STRUCTURE (GITHUB REPOSITORY)

The GitHub repository is organised in a clear and simple way so that anyone can follow the full workflow of the project. The Data/Raw folder contains the original dataset, including the file main_salary_dataset.csv.

All figures are saved in the Figures folder. This includes the year-based charts, the average salary plot, the remote-work visualisation, the country distribution chart, and the SHAP feature-importance figure. Saving these files makes it easy to view every chart without running the notebook again.

The main notebook is salary_analysis_main.ipynb, and it is stored inside the Notebook directory. It contains the full workflow from data cleaning to feature engineering, modelling, and evaluation. The requirements.txt file lists the packages used in the project so the same environment can be recreated on any machine.

6.3 CODE FLOW & EXPLANATION

The code in the main notebook follows a simple step-by-step flow so the whole process is easy to understand. First, the dataset is loaded from Google Drive and checked to make sure all columns are readable. After this, the data is cleaned by removing duplicated rows. Some new

columns are created, such as seniority level, remote category, and time-based flags. These help the model understand the structure of the data better.
When the data is ready, we are splited data into training and test sets. Numerical columns are scaled, and categorical columns are encoded, so the algorithms can work with them. Three models are trained: Linear Regression, Decision Tree, and Random Forest. Their predictions are compared to the real salary values to measure accuracy.
Finally, GridSearchCV is used to fine-tune the best model, and SHAP values help show which features have the strongest effect on the predictions.

6.4 MODEL TRAINING & PREDICTIONS

After preparing the dataset, the models were trained using the processed training data. The same preprocessing steps were applied inside the pipeline, so each model received the correct input format. Three algorithms were used in this project: Linear Regression, Decision Tree, and Random Forest. Each model learned the relationship between the features and the salary values in the training set.
Once training was finished, the models were tested on the separate test set. This allowed a fair comparison because the test data was not seen during training. Each model produced a list of predicted salary values, which were then compared with the real salaries. The Random Forest model gave the most stable results among the three, although the difference between the models was small. These predictions were later used to calculate MAE and R^2 scores, which show how well each model performs.

6.5 PERFORMANCE CALCULATION

To compare the models, the predictions from the test set were evaluated with two simple metrics: MAE and R^2. MAE shows the average difference between the real salary and the predicted value. A smaller MAE means the prediction is closer to the true number. R^2 shows how much of the salary variation the model can explain.
In this project, all three models gave similar results, but with small differences. Linear Regression reached an MAE of around 51,132 USD with an R^2 of 0.18. The Decision Tree model performed almost the same, with an MAE near 50,867 USD and an R^2 of 0.19. The Random Forest model had the best scores among the three. Its MAE was about 50,764 USD, and its R^2 reached 0.19, which was the highest.
Although the models do not predict salaries with very high accuracy, they still show clear patterns in how different factors influence salary levels.

CHAPTER 7
VALIDATION & EXTERNAL TESTING

7.1 EXTERNAL DATASET SELECTION

To test whether the final model works outside the training data, an external dataset was needed. For this purpose, a large public dataset from the Stack Overflow Developer Survey was selected. This dataset contains detailed information about professional developers, including job type, experience, working style, company size, and yearly compensation. Because it is collected independently from the Kaggle salary dataset, it allows a realistic check of how well the model performs on new data.

The original file included 84 columns and 89,184 responses. To focus on relevant cases, only full-time professional developers with a valid yearly salary were kept. After filtering, 39,623 records remained. This cleaned dataset was used to build the external test sample for evaluating the model in the following steps.

7.2 APPLYING THE FINAL MODEL TO EXTERNAL DATA

In this section of project, the final Random Forest model and the same preprocessing pipeline from Chapter 5 were applied to the external developer survey dataset. After filtering, the external data contained 39,623 rows for professional, full-time developers with a non-missing yearly salary. From this dataset, the same ten features were constructed as in the main Kaggle dataset: work year, seniority level, remote work ratio and category, employee and company country, same-country flag, post-COVID flag, post-AI flag, and company size.

These features were transformed with the fitted ColumnTransformer, and the trained Random Forest was used to generate salary predictions in USD. Two new columns add to table, a Predicted_Salary_USD and a Actual_Salary_USD columns. The first rows show that the model tends to underestimate very high salaries (for example, predicting around 55,560 USD for an observed salary of 285,000 USD), which already suggests that the external salary distribution is more extreme than the training data and will need closer analysis in the next section.

7.3 COMPARING RESULTS & ERRORS

When the final model was applied to the external dataset, the prediction errors were measured for three time ranges. These ranges were the full period (2020–2025), the post-COVID period (2021–2025), and the post-AI period (2023–2025). Since all entries in the external dataset belong to the year 2023, the same group of records was used in all three calculations. For this reason, the error values were identical across the three ranges.

The model reached a Mean Absolute Error (MAE) of about 61,369 USD on all three intervals. Between the predicted and the actual salaries, the average difference was around 27,588 USD.

Predicted model tends to underestimate the higher salaries in our external data. Even though the time-range results are the same, the comparison still confirms how the model behaves when tested on data from outside the main dataset.

7.4 DISCUSSION OF DIFFERENCES

When I compared the predictions on the external dataset with the results from the main dataset, the differences were clear. The model behaved normally on the data it was trained on, but when I used it on the new dataset, the errors became much larger. One simple reason is that the external dataset includes many developers from countries where salaries are usually high, especially the United States. Because of that, the actual salaries in the external file were much higher than the model expected, so the predictions often came out too low.
Another point is that the information in the two datasets was not recorded in the same way. Some columns had different names or formats, and after converting them, they still did not carry the same level of detail. These diferencess explain that why the model performed well inside the original dataset but struggled more when it faced new data.

7.5 RELIABILITY CHECK

In this part I tried to see how steady the model behaves. I checked the results several times, especially when I changed small things in the data. Inside the main dataset, the model stayed almost the same and did not jump too much, which gave me some confidence that it is not unstable.
Then I looked at the external dataset. Here the situation was different. The errors were higher, and the model did not follow the same pattern anymore. It seems the model handles data that looks like the training set much better than completely new cases. So, the reliability is good inside the original environment, but it becomes weaker when the data comes from another place or has very different salary levels.

CHAPTER 8
RESULTS & ANALYTICAL INSIGHTS

8.1 MODEL COMPARISON

The three machine learning models were first compared on the internal test set of the main dataset. All models were trained on the same features and evaluated with Mean Absolute Error (MAE) and the coefficient of determination (R^2). The goal was to find a model that gives reasonably low error while keeping the structure simple enough to interpret.

The linear regression model results the MAE, 51,132 USD and R^2 is around 0.19. The decision tree gave a slightly lower MAE of about 50,867 USD and a similar R^2 of roughly 0.19. The random forest performed best within this group, with an MAE of about 50,764 USD and an R^2 close to 0.19.

The differences in MAE between the three models are small, but the random forest is still the most accurate. It also tends to be more stable than a single decision tree because it averages many trees. For this reason, the random forest was selected as the main model for later interpretation and external validation.

8.2 FEATURE IMPORTANCE & SHAP INTERPRETATION

In this part, I tried to understand which features the model uses the most when it predicts salaries. I checked this in two ways. The first one was the feature importance values inside the Random Forest. The second one was SHAP, which shows how each feature pushes the prediction up or down. The full results are already shown in Table 5.1 and Figure 5.1.

Both methods pointed to the same main result:

the seniority level is the strongest feature.

When the seniority level goes up, the predicted salary also becomes higher. This matches what usually happens in real jobs, because people with more experience or responsibility normally earn more.

The next noticeable feature was the employee residence, especially the US records. In both outputs, the US code pushes the salary prediction upward. This also makes sense, since average pay in the US is generally higher than many other countries.

Other features, such as work year, remote ratio, and the type of work setting (remote, hybrid, on-site), had smaller effects. Their influence is still visible, but they do not change the salary prediction as strongly. For example, in the SHAP results, fully remote jobs slightly increase the prediction, while fully on-site jobs tend to lower it a bit.

What gave me more confidence was that the Random Forest importance values and the SHAP results followed the same structure. Both of them highlighted similar features and showed similar directions. This means the model behaves in a stable and understandable way, and its decisions can be explained clearly.

Overall, this section shows that the model mainly relies on seniority and location, while the remaining features adjust the prediction in smaller ways.

8.3 SALARY PATTERN CHANGES

To understand how salaries changed over time, I reviewed the dataset in three separate time windows. Looking at the years in smaller segments helped me see whether major global events, such as the COVID-19 period or the rise of modern AI tools, had any visible effect on salary levels.

1) Overall Trend from 2020 to 2025 (Figure 8.1)

When the whole period is viewed together, the numbers show a clear upward trend.

In 2020, the average salary was about 102,251 USD. In 2021, the average dropped slightly to 99,310 USD, which is the lowest point in the series. After that, the increase becomes steady.

In 2022 the average jumps to 131,119 USD, followed by another rise in 2023, reaching 150,118 USD.

The increase continues in 2024 and 2025, with averages of 152,312 USD and 152,824 USD, respectively. This long-term view suggests that the most important growth started after 2021 and continued in a stable way.

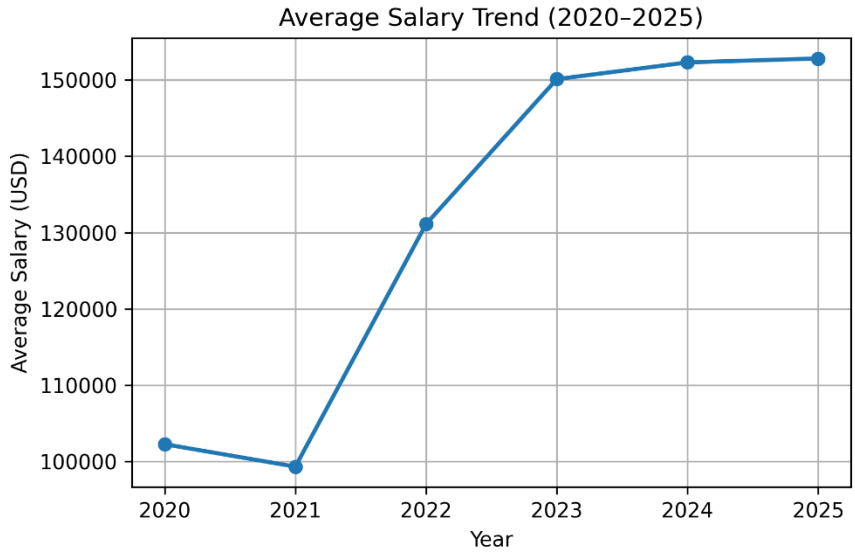


FIGURE 8.1 AVERAGE SALARY TREND FROM 2020 TO 2025

2) Post-COVID Period: 2021 to 2025 (Figure 8.2)

When I focus only on the years after COVID-19, the pattern becomes clearer.

The lowest value still appears in 2021, but as soon as 2022 begins, salaries rise sharply and then continue to grow in 2023, 2024, and 2025.

This suggests that the labour market recovered after the pandemic years, and many companies returned to normal hiring and budgeting practices, which may explain the rise in average salaries.

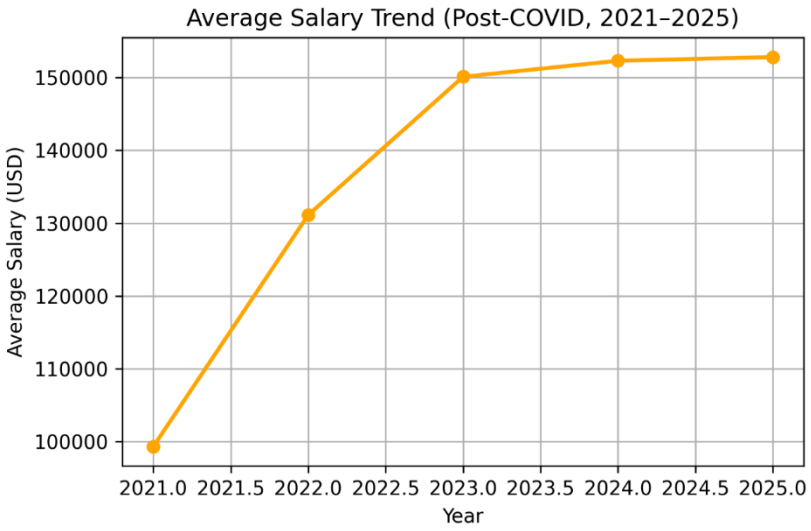


FIGURE 8.2 POST-COVID AVERAGE SALARY TREND FROM 2021 TO 2025

3) Post-AI/ChatGPT Period: 2023 to 2025 (Figure 8.3)

Looking only at the period influenced by the recent AI wave, the values remain high and fairly stable. The average in 2023 is 150,118 USD, then 152,312 USD in 2024, and 152,824 USD in 2025. The growth here is smaller compared to the previous windows, but the important point is the stability. Salaries stay at the upper level and do not show a decline.

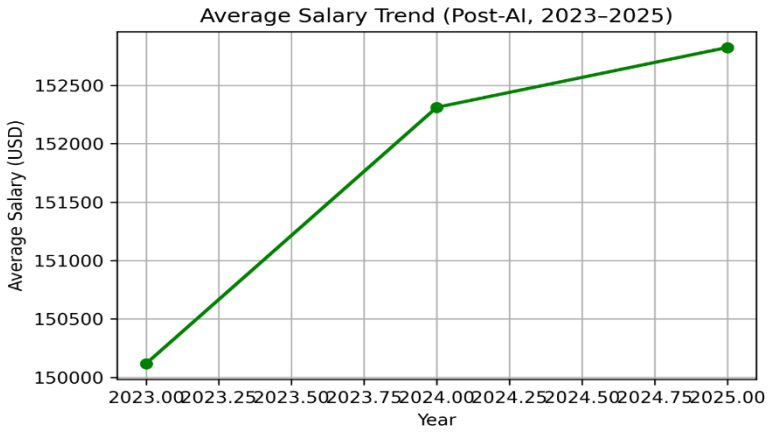


FIGURE 8.3 POST-AI/CHATGPT AVERAGE SALARY TREND FROM 2023 TO 2025

This may indicate that the introduction of AI tools did not reduce salary levels. Instead, companies still need skilled workers who can work with new technologies, which keeps salaries high.

Summary of Findings
Across all three windows, a few points stand out:

- The only drop appears in 2021, which may still reflect the uncertainty during the early COVID period.
- From 2022 onwards, salaries rise continuously.
- The AI period (2023–2025) keeps salaries high and stable rather than reducing them.
- Overall, the direction of change is positive, showing steady growth in average salary over time.

8.4 SPECIAL FOCUS ON VISUAL EVIDENCE

When I look at the salary patterns in Figures 8.1–8.3, the general direction is easy to see. Pay levels for data jobs are lowest in 2020 and 2021, and after that the numbers go up almost every year. The jump after 2022 is quite strong. Some earlier studies also mention that the early COVID-19 period created a lot of stress in the job market, but later the situation changed and digital jobs became even more important, especially when remote work became common (Tarca et al., 2024).
The SHAP results from Chapter 5 also match what we said above. People with more experience and workers living in the United States usually earn more money. This pattern appears in my model and highlights experience, country, and company features as main reasons behind salary differences, even during difficult times.
From 2023 our years overlap with the growing use of tools like ChatGPT. AI can help productivity and may create a bigger gap between workers with strong skills and those who do more routine tasks. In our data, salaries keep rising after 2023, so it seems that the spread of AI has helped experienced data professionals keep a strong position in the job market.

8.5 BUSINESS IMPLICATIONS FOR HR & COMPANIES

The model built in this project can give HR teams a clearer view of how different factors shape salaries in data-related jobs. The strong influence of experience level and country location shown in the SHAP results helps companies understand which roles need higher budgets and where salary competition is strongest. We are supporting HR team for planing the hiring decisions and salary adjustments. Thu, our model can help HR department to adjust fairly payment patterns for future planning and help maintain their employee and employers structure.

CHAPTER 9
CONCLUSION & FUTURE WORK

9.1 SUMMARY OF FINDINGS

This project followed a clear and practical goals: to understand which factors shape salaries in data-related jobs and to test how well a machine-learning model can predict these salaries. The whole analysis was built step by step. First, the dataset was cleaned and prepared, and then several new features were created, such as seniority level, remote-work category, and indicators for post-COVID and post-AI years. These steps made the dataset more suitable for modelling and helped capture the main changes in the job market between 2020 and 2025.

When the models were trained, the Random Forest model performed slightly better than the others, although overall accuracy remained modest. The scores showed that salary prediction is not a simple task because salaries depend on many personal and company-level factors that do not appear in the dataset. Even with these limits, the model was able to show which inputs matter most. Seniority level was the strongest feature, followed by the employee's country of residence. Work year, remote ratio, and some country categories also played smaller roles.

SHAP values confirmed these patterns. They showed that changes in salary predictions are mostly explained by seniority and the advantages linked to living or working in certain countries, especially the United States. This explains why some predictions stood far from the real salaries. The differences across countries and company types are large, and the dataset cannot capture everything behind them.

The external validation with the StackOverflow dataset helped check how the model behaves on unseen data. Finding results show that the errors were higher because the second dataset has a different structure and salary distribution. Still, the external dataset kept the same pattern: people with more experience and people from high-income countries receive higher predicted salaries. This means the model is consistent, even if it is not fully accurate for real-world prediction.

Finally, the salary-trend analysis from 2020 to 2025 showed a clear rise after 2021. Salaries grow strongly in 2022 and continue to increase slightly each year after that. The two special periods—post-COVID (2021 onward) and post-AI (2023 onward)—show smoother and more stable increases. This matches the overall pattern inside the dataset, where technology roles appear to gain value over time.

9.2 KEY INSIGHTS (TECHNICAL + BUSINESS)

Technically, the results from this project show that feature engineering can make a big and dramatic difference. Mapping experience between seniority levels, creating remote work groups, and adding simple indicators for the post-Covid and post-AI years helped the models read the

dataset more clearly. Even though the final accuracy remained limited, the consistency of the model across internal and external validation suggests the preprocessing pipeline was well built.
The SHAP analysis also provided practical value. Instead of only looking at accuracy numbers, SHAP helped explain why the model behaves the way it does. It showed that seniority is always the main factor shaping predictions. This is not surprising for us, but it confirms that the model focuses on real and meaningful features. Another practical insight is that the country of residence has a strong influence even after other variables are controlled. This means location still shapes salary expectations in the global tech market.
From a business point of view, these results can help HR teams and companies think about salary planning. Even if the model cannot predict exact amounts, it helps identify clusters of workers who are likely to fall into certain salary ranges. It can also show where salary differences are unusually high and help employers check whether these differences are reasonable or need attention. The clear upward trend from 2020 to 2025 also suggests that salaries in data-related fields continue to rise, especially after remote work and AI tools became mainstream.

9.3 LIMITATIONS

There were several limitations to the project that limited the accuracy of the models. The most important of these was the dataset itself. It only included information reported by individuals and did not include important variables such as the exact job role, industry subgroup, company revenue, benefit packages, bonuses, or performance levels. Salaries in real companies generally depend heavily on these elements, so the model cannot reflect the full picture.
Another limitation is the uneven distribution across countries. A large number of observations come from only a few countries, most notably the United States. This makes it difficult for the model to learn stable patterns for smaller countries. The same problem appears in foreign datasets, where the distribution of salaries has a very different shape. Because of these differences, the model's predictions for foreign data are not close to the true values.
A third limitation is that salary forecasts are inherently unstable. Two people with similar skills can receive very different offers depending on negotiations, personal networks, or hiring policies within their company. These elements cannot be captured in any data set. For this reason, even robust models will always show some error in salary predictions.
Finally, COVID-19 and AI indicators are only simple year-based indicators. They cannot capture the deeper, more gradual effects of these events. While year-to-year indicators help identify general patterns, they do not measure the precise impact of these shocks.

9.4 FUTURE RESEARCH DIRECTIONS

Future studies and works could improve the model in many different ways. First, adding more detailed features would help increase accuracy. Variables such as job function, industry, company revenue, education, skill levels, and remote-work rules would allow the model to better explain salary differences. These features maybe exist in some public datasets, and merging them with the current dataset could create a stronger foundation.

Second, applying more advanced models may help, especially gradient-boosting techniques or neural networks that are designed for complex patterns. These models may detect salary structures that Random Forests cannot capture.

Third, a more refined analysis of the COVID-19 and AI periods could lead to clearer insights. Instead of using year-based flags, future work could include month-based data or variables describing how companies changed their hiring policies.

Fourth, external validation could be expanded using datasets from Glassdoor or Indeed, where job postings include more details such as job titles and skills. This would help test how well the model generalizes in different environments.

Lastly, future work could explore fairness analysis. Since salary differences across countries and experience levels are large, checking for bias in prediction models can help employers and researchers understand whether the models create or reinforce unfair patterns.

REFERENCES

1. Ayua, S. I., Malgwi, Y. M., & Afrifa, J., (2023). 'Salary Prediction Model for Non-academic Staff Using Polynomial Regression Technique.' ***Artificial Intelligence and Applications***, 2(4), pp. *330-337*. [Online] Available at: https://doi.org/10.47852/bonviewAIA3202795 (Accessed: 13 Jun 2023).
2. *Deb, D., (2025). 'Data Science Job Salary Prediction Using Linear Regression',* ***STARS: Data Science and Data Mining***. 41. [Online] Available at: https://stars.library.ucf.edu/data-science-mining/41.
3. De la Torre-Ruiz, J.M., Cordón-Pozo, E., Vidal Salazar, M.D. and Ortiz-Perez, A. (2024) 'Pay information and employees' perception of organizational support: the mediating role of pay satisfaction', ***Employee Relations: The International Journal***, 46(9), pp. 161–177. [Online] Available at: https://doi.org/10.1108/ER-07-2023-0356 (Accessed: 05 Aug Sep 2024).
4. Goerke, L., Pannenberg, M. (2025) 'Minimum wage non-compliance: the role of co-determination', ***European Journal of Law and Economics,*** **60**, pp. 365–402. [Online] Available at: https://doi.org/10.1007/s10657-024-09811-1 (Accessed: 14 Sep 2024).
5. Gomes, N.F. and Brito, E.C. (2025) 'The impact of AI on job market dynamics and human interaction: perceptions from computer engineering students', ***Procedia Computer Science***, 270, pp. 4392–4401. [Online] Available at: https://doi.org/10.1016/j.procs.2025.09.564 (Accessed: 01 Jan 2025).
6. Juarez, L. and Villaseñor, P. (2024) 'Effects of the COVID-19 Pandemic on the Labor Market Outcomes of Women with Children in Mexico', ***Economía***, 23(1), pp. 30–49. [Online] Available at: https://www.jstor.org/stable/27366129 (Accessed: 26 November 2025).
7. Khosa, J., Mashao, D., & Subekti, F. (2024). Analyzing the Impact of Company Location, Size, and Remote Work on Entry-Level Salaries a Linear Regression Study Using Global Salary Data. ***International Journal of Informatics and Information Systems****, 7*(3), 111-122. [Online] Available at: https://doi.org/10.47738/ijiis.v7i3.215 (Accessed: 01 Sep 2024).
8. Maidin, S., Yi, D., & Ayyasy, Y. (2025). 'A Multiple Linear Regression Approach to Predicting AI Professionals' Salaries from Location and Skill Data'**.** ***International Journal of Informatics and Information Systems****,* *7*(3), pp. 100-110. [Online] Available at: https://doi.org/10.47738/ijiis.v7i3.213 (Accessed: 01 Sep 2024).
9. Ramkumar, P.N., Bernstein, J.A., Landy, D.C., DeMik, D.E., Deen, J.T., Olsen, R.J. and Cohen-Rosenblum, A. (2024) 'Determinants of Salary Variation and the Gender Pay Gap: A Survey of the American Association of Hip and Knee Surgeons (AAHKS) Surgeon Member Workforce', ***Arthroplasty Today***, 30. [Online] Available at: https://doi.org/10.1016/j.artd.2024.101554 (Accessed: 22 Sep 2024).
10. Salama, M. (2024), 'Optimization of Regression Models Using Machine Learning: A Comprehensive Study with Scikit-learn'. ***IUSRJ International Uni-Scientific Research Journal,*** 5(16), pp. 119-129. [Online] Available at: https://doi.org/10.59271/s45500.024.0624.16 (Accessed: 02 Sep 2024).
11. Semenenko, O., Sliusarenko, M., Onofriichuk, A., Onofriichuk, V. and Remez, A. (2024) 'Impact of the Russian–Ukrainian War on the National Economy of Russia', ***Journal of***

Interdisciplinary Economics, 36(1), pp. 41–57. [Online] Available at: https://doi.org/10.1177/02601079231207489 (Accessed: 01 Jan 2024).

12. Tarca, V., Luca, F.-A., & Țarca, E. (2024) 'The Digital Edge: Skills That Matter in the European Labour Market after COVID-19', ***Economies***, *12*(10), 273. [Online] https://doi.org/10.3390/economies12100273 (Accessed: 08 Oct 2024).
13. Zarifhonarvar, A. (2023) 'Economics of ChatGPT: a labor market view on the occupational impact of artificial intelligence'**, *Journal of Electronic Business & Digital Economics***, 3(2), pp. 100–116. [Online] Available at: https://doi.org/10.1108/JEBDE-10-2023-0021 (Accessed: 14 Nov 2023).

APPENDICES

APPENDIX A: GITHUB REPOSITORY

https://github.com/behrad-mohammadi/salary-prediction-data-jobs.git

APPENDIX B: MAIN DATASET ADDRESS LINK

https://www.kaggle.com/datasets/saurabhbadole/latest-data-science-job-salaries-2024

APPENDIX C: EXTERNAL DATASET ADDRESS LINK

https://www.kaggle.com/datasets/mehranmahdiani/stackoverflow-developer-survey-2023/data