DATA SCIENCE SALARY PREDICTIONS MACHINE LEARNING PROJECT

Angelina Aziz

Loughborough University
22WSP774 Machine Learning – Principles & Applications for Engineers

TABLE OF CONTENTS

TABLE OF FIGURES

ABSTRACT

As the field of data science continues to expand, the demand for skilled professionals in this area has increased dramatically. To make informed decisions about hiring and compensation, it is essential to understand the market trends and factors influencing data science salaries. This paper presents an end-to-end machine learning project for predicting salaries in the data science industry using the Data Science Salary dataset [1]. The project involves several steps, including data cleaning, exploratory data analysis and modelling. We use various machine learning algorithms, including polynomial regression, random forest, and XGBoost, to predict the salary of data science professionals. Additionally, we implement a stacking ensemble model, which combines the predictions of multiple models, the best performing model was found to be a XGBoost model.

Our approach highlights the importance of data cleaning and model selection in improving prediction accuracy. Additionally, the project provides valuable insights into the factors influencing data science salaries, including job title, years of experience, and location. Overall, this project demonstrates the potential of machine learning techniques in predicting salaries in the data science industry and the value of end-to-end projects in providing a comprehensive understanding of the machine learning pipeline. As the demand for skilled data scientists continues to rise, accurate salary predictions can aid in the hiring process, benefit both job seekers and employers, and help to promote a fair and equitable job market.

INTRODUCTION

In the age of data-driven decision-making, predicting salary trends and understanding the factors that influence them has become a vital aspect of the recruitment industry. With the increasing demand for data scientists, there is a growing need to accurately predict their salaries based on their skills and experience. To address this need, we turn to the Data Science Salary dataset provided by Kaggle, which offers a valuable opportunity to develop and test machine learning (ML) models that can accurately predict salaries for data scientists based on their skills and experience.

In this paper, we present an end-to-end ML project that explores various techniques and algorithms to predict salaries for data scientists from the Data Science Salary dataset. We begin by performing exploratory data analysis (EDA) to gain insights into the factors that influence data science salaries and identify trends in the data. Next, we pre-process and clean the data, applying various feature engineering techniques to extract relevant features that can improve the accuracy of the model. We then train and evaluate several ML models, including regression, decision trees, random forests, and gradient boosting, to identify the most accurate and efficient model for predicting data science salaries.

Overall, our project demonstrates the value of ML in predicting salaries for data scientists and provides practical insights and techniques that can be applied to similar datasets in the future. By leveraging the power of ML, recruiters can gain a deeper understanding of the factors that influence data science salaries and make data-driven decisions that can help attract and retain top talent in the field.

DATA SELECTION & OBJECTIVES

The selection of an appropriate dataset is a critical step in the success of any machine learning project [2]. In this section of the report, we explain the rationale behind choosing the Data Science Salary Prediction dataset and outline the benefits and challenges associated with its use.

The Kaggle Data Science Salary dataset contains salary information from various data science roles worldwide. The dataset includes information such as job title, years of experience, education level, company size, and compensation. The goal of the dataset is to provide insights into the factors that influence data science salaries and to help professionals make informed decisions about their careers. One of the main reasons we chose this dataset is that it offers valuable insights into the data science job market and is relevant to current trends in the industry. As data science continues to grow in popularity and demand, understanding the factors that affect salaries is crucial for professionals and employers alike. The dataset is also substantial in size, with over 3,000 entries and 11 features, providing a comprehensive and diverse set of data for analysis.

One of the main reasons this data set was chosen is that it provides a realistic and challenging problem that is relevant to real-world applications. Data science is a rapidly growing industry, and predicting fair market salary is crucial for recruiters and data scientists alike.

BENEFITS OF CHOSEN DATA

There are several benefits of using the selected data science salary dataset. Firstly, it provides a large-scale dataset with diverse features, including education level, years of experience, and job title, which can help to identify key factors that impact salaries in the data science industry. Secondly, working with a real-world dataset provides an opportunity to develop practical skills in data cleaning, exploratory data analysis, and feature engineering, which are crucial for success in real-world data science applications.

The dataset has been sourced from aijobs.net an online job board specifically focused on jobs in the AI space and is proven to be legitimate non-generated data. This means that the data should be highly representative of the real data science job market and can be treated as such.

Additionally, the dataset contains a wide range of salary levels across different geographic locations and industries, which can provide valuable insights into the job market and help individuals make informed decisions about their careers. Another benefit of using this dataset is its potential real-world applications. Companies can use this type of data to gain a better understanding of the job market and make informed decisions about compensation packages, which can lead to increased employee satisfaction and retention.

CHALLENGES OF CHOSEN DATA

There are several challenges associated with working with the data science salary dataset. One of the challenges is the presence of a mix of categorical and numerical data which can make it difficult to identify correlations between features. Another challenge is the need for careful data pre-processing and feature analysis to understand the significance of certain features. The dataset also contains a relatively small number of observations (fewer than 4,000 rows) and limited features (only 11 fields), which may limit the scope and accuracy of the models produced.

Finally, the dataset may not account for external factors that may impact salary decisions, such as gender, age, cost of living, and level of education, which may limit the generalizability of the insights and models produced.

OBJECTIVE AND MOTIVATIONS

The main objective of analysing the Data Science Salary dataset is to gain insights into the factors that influence the salaries of data science professional and create a salary prediction model. The motivation for this project is to better understand the job market for data scientists and to identify trends that can help professionals negotiate their salaries and plan their careers. Additionally, the project aims to provide insights to companies on the factors that are most important in determining the salaries of data science professionals, which can help them make more informed hiring and compensation decisions. Ultimately, the goal is to create a predictive model that accurately estimates data science salaries based on relevant factors such as job title, experience etc.

DATA PREPERATION & ANALYSIS

The term data analysis refers to the processing of data by conventional theories, technologies and tools for extracting useful information for practical purposes [2] In this case we must prepare and clean the data before we can process it. We will first check the dataset for null values and duplicates before analysing each individual feature before analysing key features against the target feature,

THE DATA AT FIRST GLANCE

The dataset under consideration encompasses 3755 instances and 11 attributes, among which 4 are quantitative, namely work_year, salary_in_usd, salary, and remote_ratio, while the remaining 7 are categorical, namely experience_level, employment_type, job_title, employee_residense, company_location, salary_currency and company_size. The data was collected over a 3-year time frame spanning from 2020 to 2023, and notably, it is free of missing values. However, it is worth noting that due to the categorical nature of most attributes, identifying duplicates without a unique identifier column is so we have disregarded duplication as the majority of duplicates identified are not true duplicates. We have also made the choice to disregard the salary and salary_currency attributes as they are already encompassed by the salary_in_usd attribute.

DATA ANALYSIS OF CATEGORICAL FEATURES

In this section, we aim to perform an in-depth exploration of the categorical features present in the dataset. We seek to identify any significant trends or patterns that may exist within the data and extract noteworthy insights. This approach will allow us to gain a comprehensive understanding of the key characteristics of the dataset and provide a solid foundation for any subsequent analysis or modelling. By examining the various categorical features, such as experience level, employment type, job title, employee residence, company location, and company size, we can uncover valuable information regarding the prevailing trends in the field of data science, as well as identify areas of interest for further research and investigation. Through this analysis, we can provide a more nuanced and detailed picture of the data, enabling us to draw more robust conclusions and make more informed decisions.

EXPERIENCE LEVEL

First, we will explore the experience level feature. There are 4 categorical values:

- EN, which refers to Entry-level / Junior
- MI, which refers to Mid-level / Intermediate.
- SE, which refers to Senior-level / Expert.

- EX, which refers to Executive-level / Director.

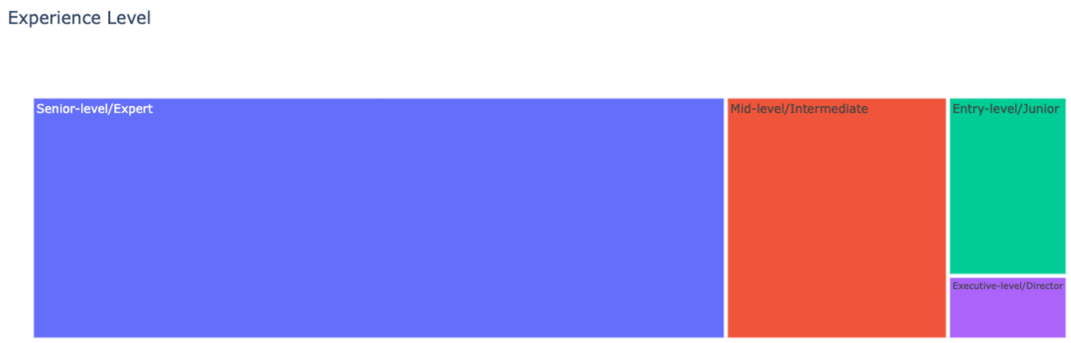


Figure 1 - Experience Level Tree Map

From the tree map above, we observe that Senior-level/Expert accounts for the highest number of entries at 2516, and Mid-level/Intermediate ranked next at 805, followed by Entry Level at 320 and Executive at the end with 114 entries.

JOB TITLES

The next feature to delve into is job title. There are a total of 93 different job titles within this dataset. Figure 2 shows a bar plot of the top 15 job titles. Data Engineers appear to be leading the job title category at 1040, followed by Data Scientist at 840 and Data Analyst at 612, at the bottom of the top 15 we have Data Analytics Manager at 22 entries.

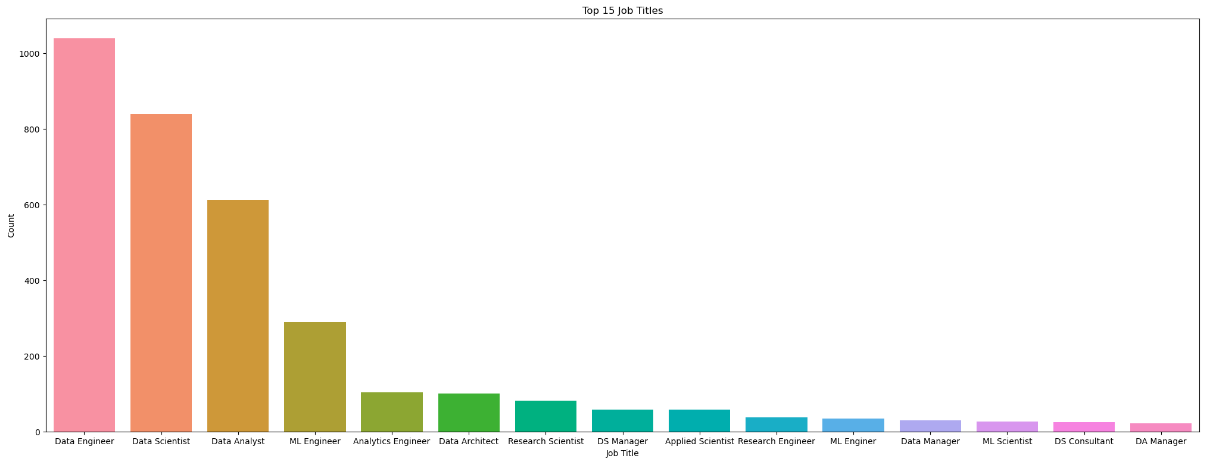


Figure 2 - Top 15 Job Titles Bar Chart

EMPLOYMENT TYPE

The next feature we will be analysing is employment type. There are 4 categorical values:

- PT: Part-time
- FT: Full-time

- CT: Contract
- FL: Freelance

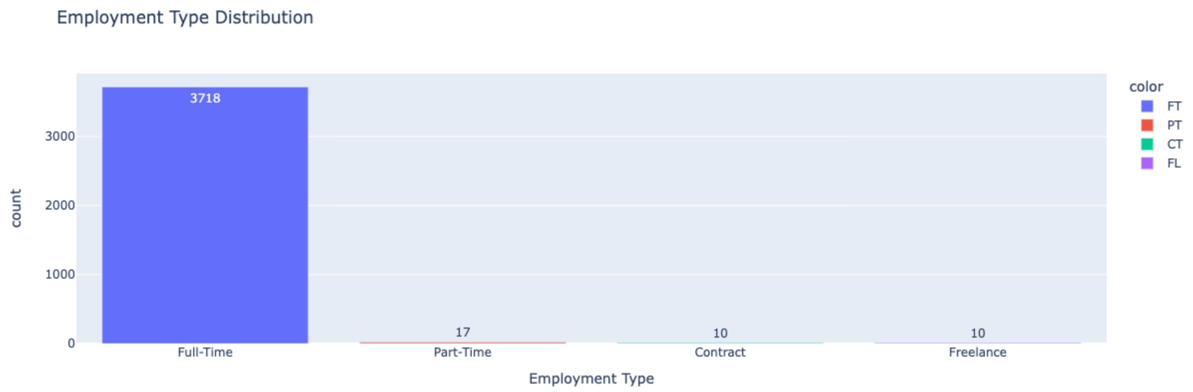


Figure 3 - Employment Type Bar Chart

It is clear from figure 3 that almost the entirety of the dataset is made up of full-time employees and there is very little values for the other 3 categories.

EMPLOYEE RESIDENCE

The dataset comprises of 78 unique countries of employee residence. Figure 4 depicts a bar plot exhibiting the top 15 countries of employee residence, wherein the United States, with 3004 employees, dominates the chart followed by Great Britain with 167 employees and Canada with 85 employees. The bottom-most positions in the top 15 countries are occupied by Pakistan and Italy, each with 8 employees.

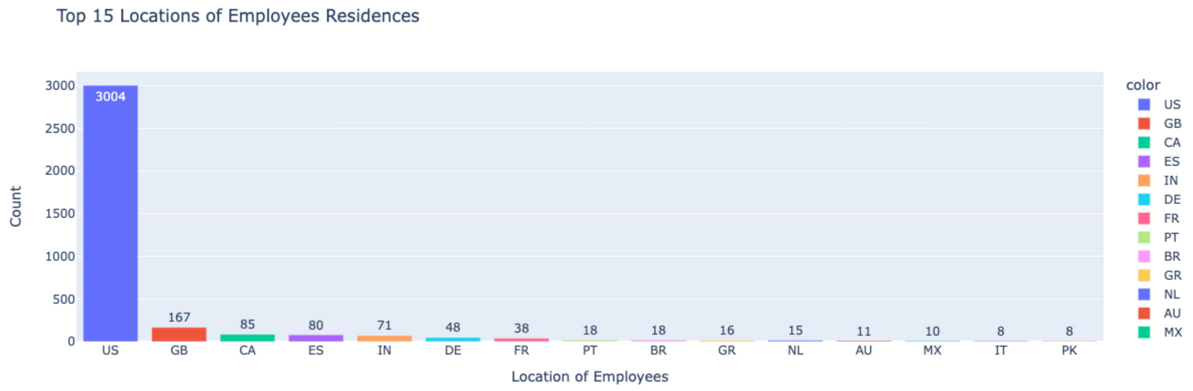


Figure 4 - Employee Location Bar Chart

To better understand the geographic distribution and extent of employee location, employee residence was visualized on a choropleth map, as shown in Figure 5.

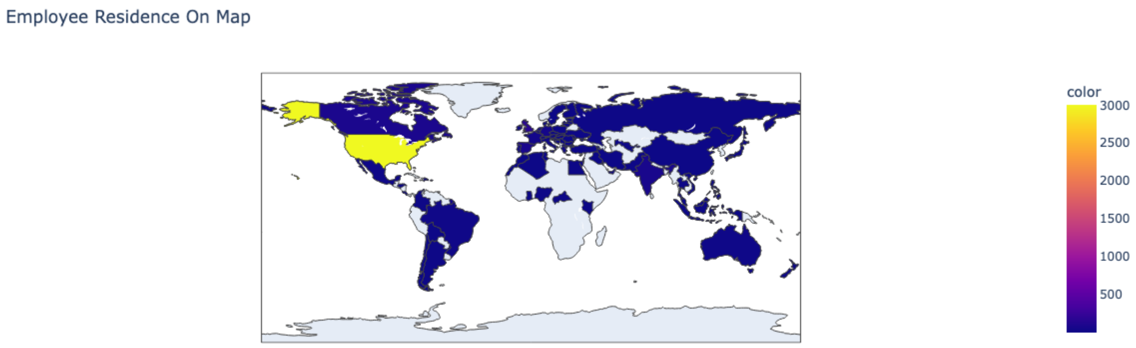


Figure 5 - Employee Residence on Choropleth Map

Both Figure 4 and Figure 5 distinctly portray that the bulk of employees in this dataset reside in the United States.

COMPANY LOCATION

The Data Science Salary 2023 dataset encompasses companies operating in 72 distinct countries. Figure 6 illustrates the top 15 countries, by count of companies, with the United States retaining a significant lead with 3040 companies, trailed by Great Britain at 172 and Canada at 87. Conversely, Ireland and Singapore occupy the bottom ranks with only 7 and 6 companies, respectively.

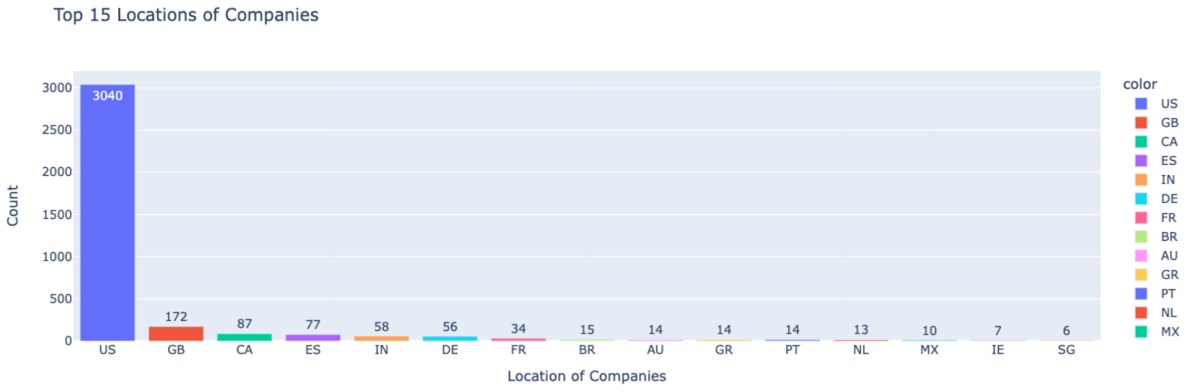


Figure 6 - Company Locations Bar Chart

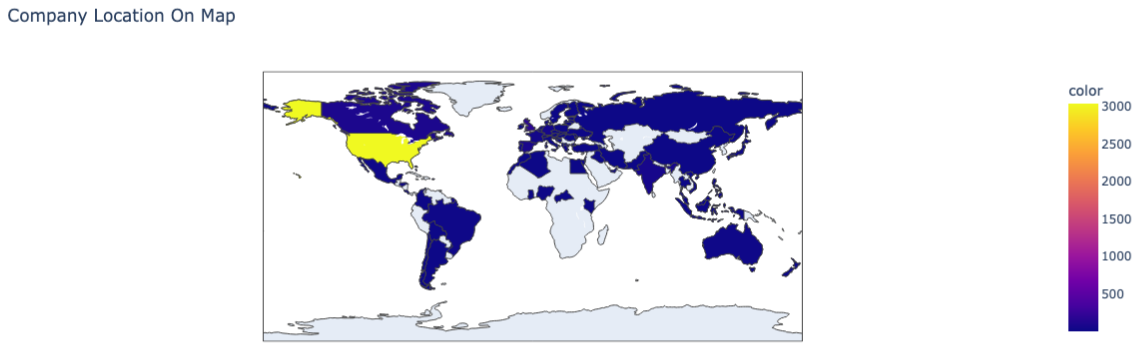


Figure 7 - Company Location on Choropleth Map

A choropleth map was utilized to visualize the geographic distribution and extent of company location, as shown in Figure 7. Both Figure 6 and Figure 7 effectively portray the preponderance of companies in this dataset that are situated in the United States.

EMPLOYEE RESIDENCE AND COMPANY LOCATION

Considering that both the employee residence and company location features are geographically related, it could prove insightful to establish a correlation between them. It is plausible to assume that the country of residence of an employee would align with the geographical location of the company they work for; however, with the growing prevalence of remote work, this assumption may not hold true. Consequently, there is value in mapping these features and examining the potential relationship between them.

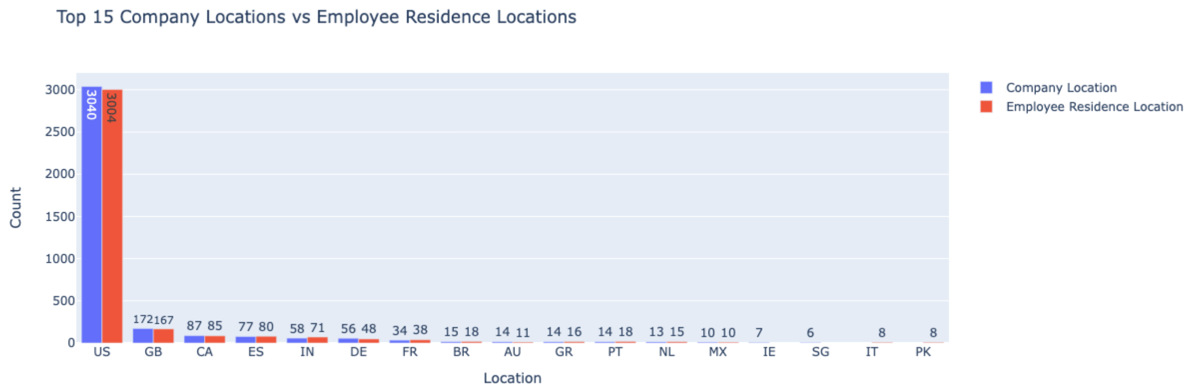


Figure 8 - Top 15 Company Locations Mapped Against Employee Residence Locations

Based on the data represented in Figure 8, which is a bar chart displaying employee residences and company locations for each country in the dataset, it is evident that a general trend exists across most countries, where the number of employee residences and company locations are similar. The absence of a clear deviation from a linear relationship between employee residence and company location indicates that employees typically reside in the countries where their respective companies are located. However, the presence of outliers in the chart suggests that external factors, such as remote work or international job assignments, may impact the geographic distribution of employees in relation to their respective companies.

COMPANY SIZE

The Data Science Salary 2023 dataset includes companies categorized into three different size classes, namely small, medium, and large. The classification criteria adopted by aijobs.net consider companies with less than 50 employees as small, those with between 50 and 250 employees as medium, and those with more than 250 employees as large. Based on the information presented in Figure 9, it is evident that the majority of companies in the dataset belong to the medium size category, with a total of 3153 respondents. Large companies rank second, with 454 respondents, followed by small companies with the lowest number of respondents at 148.

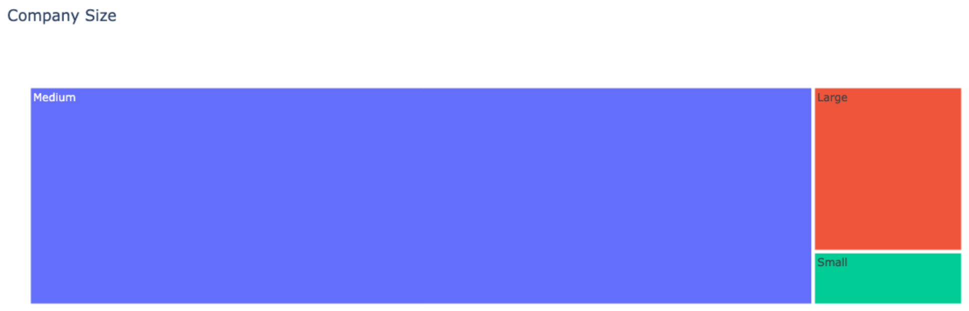


Figure 9 - Tree Map of Company Size

DATA ANALYSIS OF NUMERICAL FEATURES

In this section, we aim to conduct a comprehensive exploration of the numerical features present in the dataset. Our goal is to identify any significant patterns or trends that may exist within the data and extract noteworthy insights. This approach will enable us to gain a thorough understanding of the key characteristics of the dataset and establish a solid basis for any subsequent analysis or modelling. By examining the various numerical features, such as salary, years of experience, and age, we can uncover valuable information regarding the prevailing trends in the field of data science and identify areas of interest for further research and investigation. Through this analysis, we can present a more nuanced and detailed view of the data, enabling us to draw more robust conclusions and make more informed decisions.

WORK YEAR

Although technically a numerical feature the nature of the work year feature would be better suited to be a categorical value as there are only a select number of years the data can be from as ai.jobs started collecting this data in 2020 till the current year 2023.

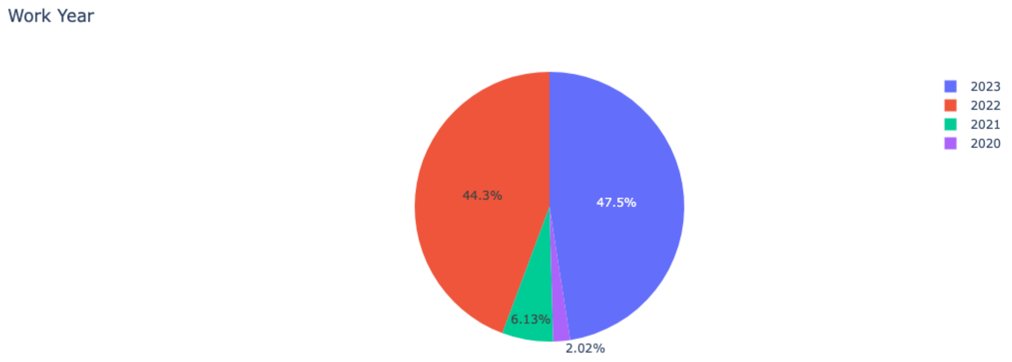


Figure 10 - Work Year Pie Chart

Figure 10 shows the dataset exhibits a notable trend in which the majority of data has been collected during the years 2022 and 2023, despite only containing the data for the first four months of 2023. This trend could be attributed to the rapid growth of the data science field over the past few years, or potentially due to increased popularity of the ai.jobs site during this time period.

REMOTE RATIO

The feature remote ratio consists of 3 values 0, 50 and 100, these can be broken down as such:

- 0: No remote work (less than 20%)
- 50: Partially remote
- 100: Fully remote (more than 80%)

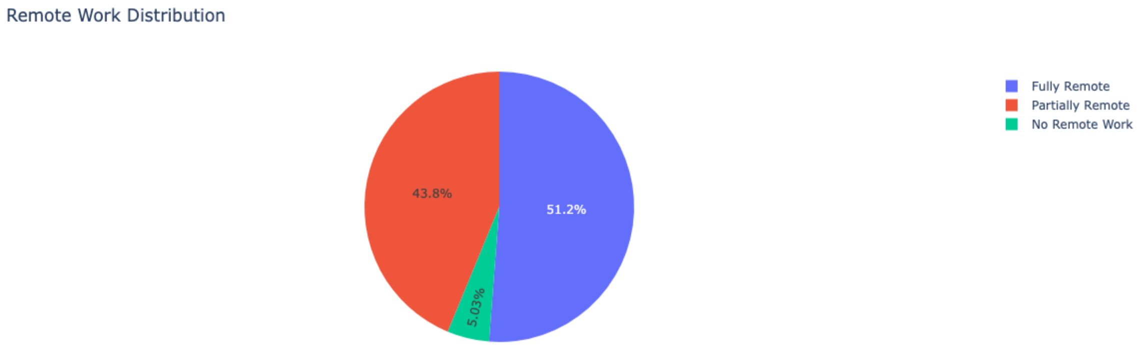


Figure 11 - Remote Work Pie Chart

From Figure 11 it is clear that the majority of employees/companies in the dataset have at least a partially remote work policy with fully remote being the most common remote work policy.

SALARY IN USD

The target variable for our prediction models is the salary in USD, a numerical feature that exhibits the highest degree of variability in the dataset. A visual representation of the salary distribution is depicted in Figure 12, which presents a box plot.

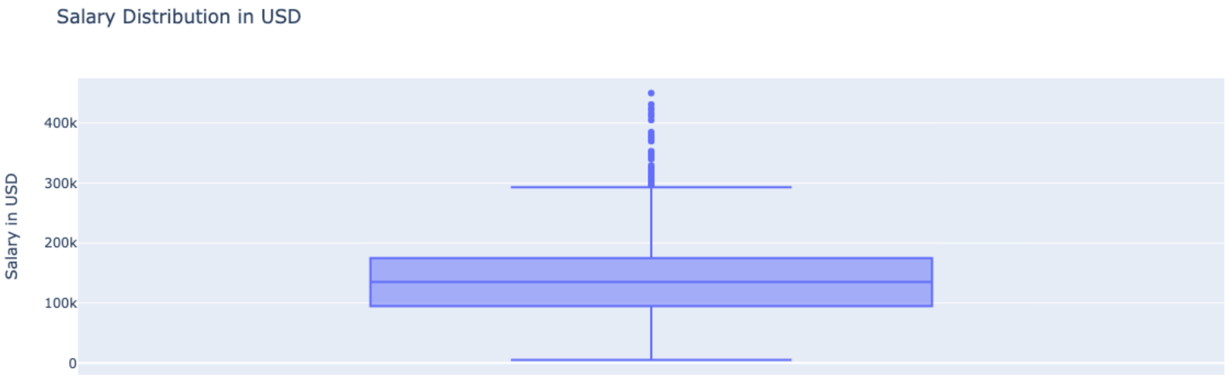


Figure 12 - Box Plot of Salary Distribution

The maximum salary in the dataset is $450k, the minimum is $5132, the majority of the salaries are between 95k and 175k with a median salary of 135k.

DATA CLEANING STRATEGY BASED ON INITIAL ANALYSIS

Given the substantial concentration of data points pertaining to employees and companies located in the United States, the dataset lacks sufficient representation of other geographical locations, thereby impeding the accuracy of data science salary predictions for employees residing outside the US. Consequently, as a strategy to enhance the reliability and applicability of predictive modelling, it is deemed necessary to exclude data points associated with non-US employee residence from the dataset. By adopting this approach, we aim to focus our analysis on a more homogeneous subset of data that aligns with the predominant geographic composition of the dataset, thereby facilitating more accurate and meaningful predictions for data science salaries.

Upon excluding non-US employees and company locations, the dataset has been reduced to 2999 entries. Following this, we aim to eliminate all non-full-time employees from the pre-cleansed dataset, as only 37 non-full-time employees are present. After this process, the cleansed dataset is left with 2991 entries. Given the initial small size of the dataset, no further cleaning will be performed to avoid potential loss of information, which may compromise the accuracy of the prediction model.

DATA ANALYSIS ON MULTIPLE VARIABLES

In this section of the report, we aim to conduct a comprehensive analysis of multiple attributes or features present in the dataset. By examining the relationships between various features, we seek to identify significant trends and patterns that can provide valuable insights into the underlying dynamics of the data science field. Through this approach, we can uncover hidden relationships and dependencies that can aid in the development of more accurate predictive models and inform decision-making processes. Furthermore, the analysis of multiple features can enable us to gain a more holistic understanding of the data, and thus facilitate the formulation of more nuanced and sophisticated interpretations of the results. Overall, this approach can provide a powerful framework for identifying key insights and unlocking new opportunities for further research and exploration.

REMOTE RATIO BY WORK YEAR

Given the unprecedented Covid-19 pandemic, it is reasonable to assume that the year in which the data was collected may have had an impact on the prevalence of remote work. In order to investigate this potential relationship, a polar scatter graph was generated, as depicted in Figure 13, to visualize the distribution of remote work across the years of the dataset.

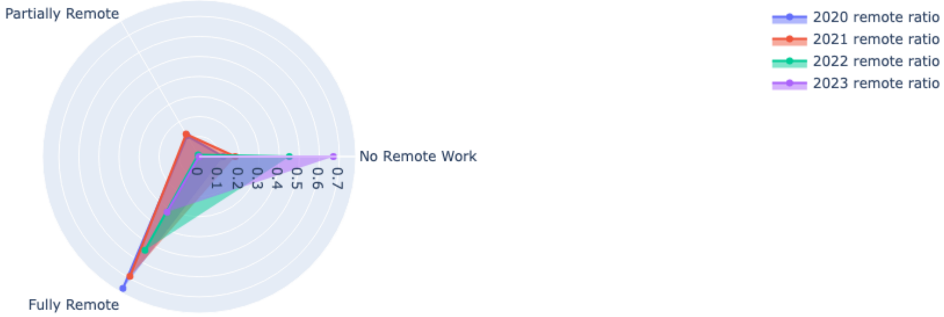


Figure 13 - Polar Scatter Plot of Remote Work by Year

The polar plot indicates that there exists a discernible pattern with regards to remote work distribution over the years. Specifically, the plot suggests that there was a peak in remote work during the year 2021, which can be reasonably attributed to the Covid-19 pandemic and its impact on the work environment. However, as the plot indicates, there has been a noticeable decrease in remote work policies in subsequent years, such as in 2023 where there is less apparent implementation of partial and full remote work policies.

SALARY BY WORK YEAR

The subsequent predictive models will primarily focus on the target feature of salary. Therefore, it is imperative to explore and scrutinize the relationships between other features and salary to reveal any significant correlations. This approach will enable us to identify potential predictors of salary, which can enhance the accuracy of our predictive models. By analysing the interrelationships between salary and work year, these insights can then be used to develop more precise models that can better predict salary levels based on the given input features.

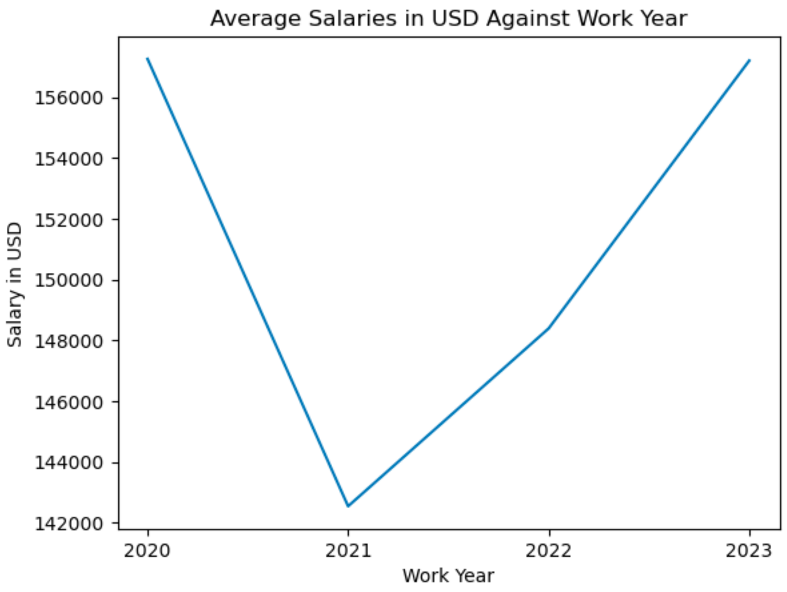


Figure 14 - Line Chart of Salary in USD Against Work Year

Upon analysis of the salary distribution, it is discernible that the average salary of data science professionals in the US experienced a decline during the year 2021, with a subsequent rebound to pre-pandemic levels in the year 2023. This trend indicates a temporal correlation between the COVID-19 pandemic or other possible factors and the fluctuation of salaries in the data science field. The aforementioned observation serves to emphasize the significance of considering temporal factors when analysing salary data. By acknowledging these fluctuations and identifying the underlying factors that contribute to such trends, we can develop more accurate predictive models that account for the nuances of the industry.

SALARY BY EXPERIENCE LEVEL

As an extension of the analysis to examine the correlation between various features and the target feature of salary, Figure 15 displays the mean salary in USD based on the experience level of data science professionals. With the Entry-Level/Junior average salary at $106.5k, Mid-Level/Intermediate at $130k, Senior-Level at $159k and Executive-Level/Director at $206k

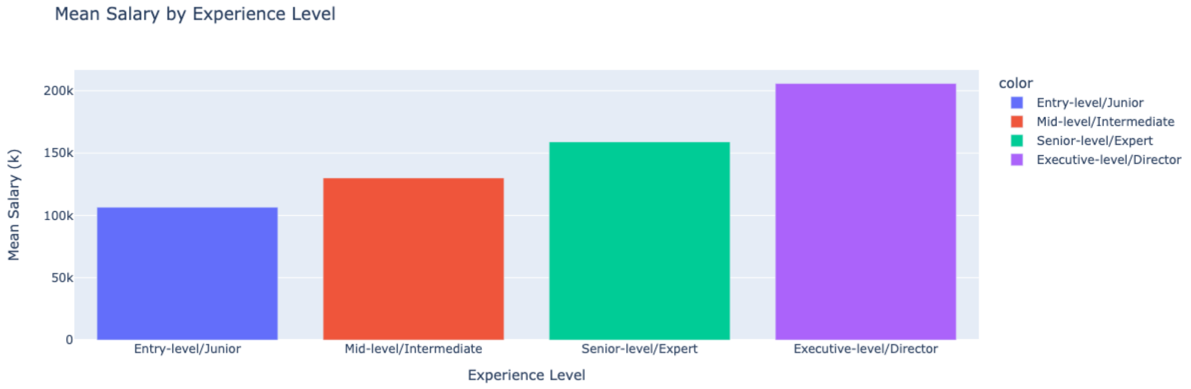


Figure 15 - Histogram of Mean Salary by Experience Level

As depicted in Figure 15, there exists a noticeable trend where individuals with a higher experience level in the field of data science tend to receive a higher average salary in comparison to their junior counterparts. This pattern is in line with expectations within the industry.

SALARY BY COMPANY SIZE

As a continuation of the analysis to investigate the relationship between various attributes and the target feature of salary, Figure 16 illustrates the average salary in USD based on the size of the company where data science professionals work. Average salary at a small company is $118k, medium is $152.8k and large companies are 158.6k

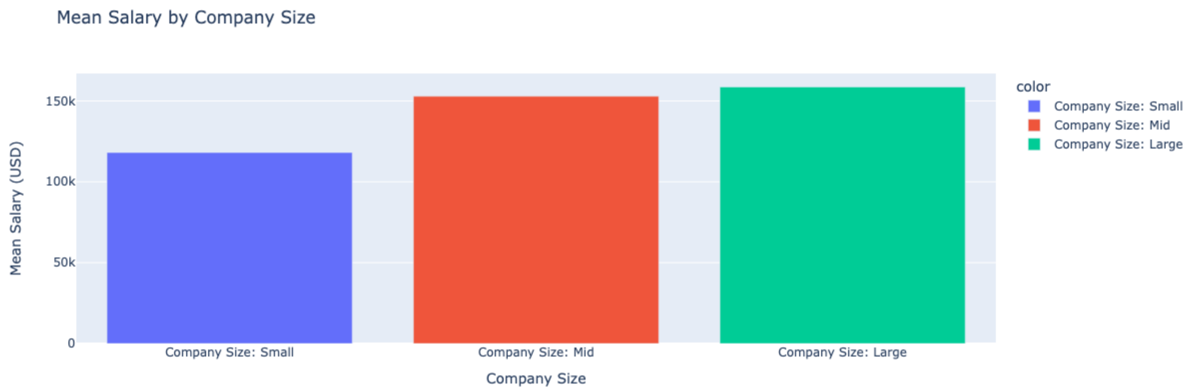


Figure 16 – Histogram of Mean Salary by Company Size

Based on the analysis of the dataset, a positive correlation was observed between the mean salary in USD and the size of the employing company, as illustrated in Figure 16. It can be inferred that as the company size increases, so does the average salary of data science professionals.

The dataset under consideration encompasses 3755 instances and 11 attributes, among which 4 are quantitative, namely work_year, salary_in_usd, salary, and remote_ratio, while the remaining 7 are categorical, namely experience_level, employment_type, job_title, employee_residense, company_location, salary_currency and company_size. The data was collected over a 3-year time frame spanning from 2020 to 2023, and notably, it is free of missing values. However, it is worth noting that due to the categorical nature of most attributes, identifying duplicates without a unique identifier column is so we have disregarded duplication as the majority of duplicates identified are not true duplicates. We have also made the choice to disregard the salary and salary_currency attributes as they are already encompassed by the salary_in_usd attribute.

FINAL FEATURE SELECTION AND CLEANING

After conducting a thorough analysis of the dataset and preparing it for model selection and training, we have determined that three numerical attributes, including work year, salary in USD, and remote ratio, along with 4 categorical attributes, including experience level, employment type, job title and company size, will be carried forward for further analysis however they will be converted to numerical attributes to be easier to work with and scale.

Furthermore, we have ensured the reliability of our dataset by removing all non-US employees and companies, as well as any non-full-time employees, resulting in a final dataset of 2991 values out of the original 3755 values. Employee residence and company location will be dropped as there is no longer a need to carry these columns around as they are now one value and offer no further insight. These steps have been taken to ensure the quality of the dataset and to increase the accuracy of our subsequent analysis and modelling.

MODEL SELECTION & TRAINING

The subsequent section of this report delves into the crucial aspect of model selection and training. This section aims to provide an in-depth analysis of various factors that were considered while choosing a suitable model, the models that were ultimately selected, the reasons behind their selection, the techniques employed for model training, and the iterative methods employed for enhancing the accuracy of the final prediction model.

MODELS CONSIDERED

LINEAR REGRESSION

Linear Regression was the first model that was considered as it is a simple and fast model to train, making it easy to understand. It performs well when there is a linear relationship between the features and the target salary feature. However, it may not perform well when the relationship is non-linear, as it assumes a linear relationship, which could be an oversimplification of the relationships between certain features in the dataset and the salary in USD feature.

RIDGE REGRESSION

Ridge Regression was subsequently evaluated as a potential model due to its ability to handle multicollinearity and prevent overfitting through the incorporation of a regularization term in the cost function. Despite its potential limitations in cases where the number of features exceeds the number of samples, such concerns were deemed inconsequential for the present dataset, given that the number of features was significantly smaller than the number of samples available for analysis.

LASSO REGRESSION

The suitability of Lasso Regression as a model for predicting data science salaries was investigated, given its ability to perform feature selection by shrinking the coefficients of less important features to zero, which reduces the complexity of the model. Nevertheless, caution must be taken when dealing with highly correlated features, as it may arbitrarily select one of the correlated features such as experience level, leading to a reduction in the model's performance.

DECISION TREES

Decision Trees were considered as a potential model for the data science salary prediction task as they are non-parametric and can handle non-linear relationships. Furthermore, they are easy to interpret and understand, which is desirable in this context. However, there is a risk of overfitting the data, which may result in poor generalization to new, unseen data. Therefore, care must be taken to prevent overfitting during the training process.

RANDOM FOREST

Random Forest is a viable candidate model for the prediction of salaries in the data science domain given its ability to handle non-linear relationships and high-dimensional data. Its ensemble learning approach helps to mitigate overfitting. However, the model's effectiveness may be impacted by features with a high degree of correlation or imbalanced data, hence the significance of meticulous feature selection and diligent performance evaluation with suitable metrics.

GRADIENT BOOSTING

The Gradient Boosting model is a suitable choice for the data science salary prediction cleaned dataset, which comprises of just under 3,000 data points and 9 features. This model can effectively handle non-linear relationships and high dimensional data, and incrementally builds the prediction model by emphasizing poorly predicted samples from previous iterations, leading to improved predictive accuracy. Although the model's complexity is relatively low, the selection of hyperparameters can significantly affect performance and may lead to high computational costs. Therefore, it is important to conduct extensive experimentation and meticulous evaluation of the model's performance, leveraging appropriate metrics.

NEURAL NETWORKS

In the context of the data science salary prediction dataset, Neural Networks can be regarded as a promising model due to their capacity to handle non-linear relationships and high dimensional data. This model's ability to automatically learn features from the data without the need for manual feature engineering makes it particularly appealing. However, due to the limited size of the dataset, Neural Networks may require meticulous hyperparameter tuning and regularization techniques to prevent overfitting. Moreover, the inherent complexity of Neural Networks can make the interpretation of results challenging. Therefore, conducting rigorous experimentation and meticulous evaluation of the model's performance using appropriate metrics is crucial.

SUPPORT VECTOR REGRESSION

SVR is a plausible choice for the data science salary prediction dataset due to its proficiency in handling non-linear relationships and high dimensional data, while performing well in high-dimensional spaces. However, hyperparameter selection and computational complexity can affect performance, even with a small dataset. Appropriate feature scaling is necessary as SVR requires scaling of the data. Meticulous experimentation and evaluation of the model's performance using appropriate metrics are crucial when considering SVR as a modelling technique for this dataset.

POLYNOMIAL REGRESSION

Polynomial regression can be a suitable modelling technique for the data science salary prediction dataset due to its ability to capture non-linear relationships between variables. However, it may overfit the data and may require careful selection of hyperparameters to avoid this issue. Additionally, polynomial regression may become computationally expensive as the degree of the polynomial increases.

ENSEMBLE MODELS

Ensemble models offer several advantages for the data science salary prediction dataset, such as combining the strengths of multiple models to potentially improve accuracy and generalization performance. They can also be resilient to overfitting if individual models are not overfitting themselves and allow for a flexible range of modelling approaches. However, there are also drawbacks, such as the computational expense of training and combining multiple models and sensitivity to model selection and hyperparameters. The suitability of ensemble modelling for this dataset ultimately depends on the dataset's characteristics and modelling objectives.

MODELS SELECTED

In the process of model selection for the data science salary prediction task, several factors were taken into consideration. Among these factors were the efficiency of model development and training, interpretability of the model, and the appropriateness of the model for the data science salary prediction dataset of 2023. The objective was to identify models that could accurately predict salary while also providing a reasonable level of transparency and ease of use.

LINEAR REGRESSION

Linear regression is a widely used and well-established modelling technique that has been applied to numerous data science problems, including salary prediction. In the context of the data science salary prediction dataset, linear regression was selected as one of the modelling techniques due to its simplicity, interpretability, and ability to handle continuous and categorical features.

According to [3], linear regression is one of the most common statistical models used in data science and machine learning applications. Linear regression models are widely used due to their simplicity and interpretability, making them an ideal choice for use in situations where the goal is to understand the relationship between a dependent variable and one or more independent variables.

In addition to its simplicity and interpretability, linear regression has also been shown to perform well on a wide range of datasets. In fact, in a study comparing various modelling techniques on a salary prediction dataset, linear regression was found to be one a great starting point for polynomial models [4]. The study also found that linear regression was able to identify the most important features for predicting salary, which can be useful for understanding the factors that contribute to salary levels in the data science industry.

Overall, the simplicity, interpretability, and performance of linear regression make it a suitable modelling technique for the data science salary prediction dataset.

POLYNOMIAL REGRESSION

Polynomial regression was selected as a modelling technique for the data science salary prediction dataset owing to its capacity to capture non-linear relationships between the target variable and features. As the association between salary and predictors like experience may not be strictly linear, polynomial regression can provide a more precise and adaptable model.

Furthermore, polynomial regression is a relatively straightforward and interpretable model compared to other non-linear models, such as neural networks. The model can be easily visualized to comprehend the relationship between predictors and the target variable.

RANDOM FOREST

Random Forest is a popular ensemble learning algorithm that can be used for regression and classification tasks. It has been applied to numerous machine learning problems and has shown to perform well on a variety of datasets including small datasets such as the data science salary prediction. In the context of the data science salary prediction dataset, Random Forest was chosen as one of the modelling techniques due to its ability to handle non-linear relationships between features and the target variable, as well as its ability to handle large datasets.

According to Breiman [6], Random Forest is a versatile machine learning algorithm that combines the strengths of decision trees and ensemble learning. Random Forest can handle a large number of input variables and can identify the most important features for predicting the target variable. It can also handle missing data and outliers, making it a robust algorithm for data analysis.

Overall, the ability of Random Forest to handle non-linear relationships, outliers and its strong performance on a variety of datasets, make it a suitable modelling technique for the data science salary prediction dataset.

GRADIENT BOOSTING (XGBOOST)

Gradient boosting, specifically XGBoost, was selected as one of the modelling techniques for the data science salary prediction dataset due to its ability to handle complex, non-linear relationships and its proven success in various machine learning applications. gradient boosting is a powerful ensemble learning method that combines weak learners, such as decision trees, to create a stronger model with high predictive accuracy [7].

XGBoost, in particular, has gained popularity in recent years due to its scalability, speed, and ability to handle both numerical and categorical data. Despite its success, XGBoost has some potential drawbacks, including the risk of overfitting, sensitivity to hyperparameters, and computational expense. However, these issues can be addressed with proper parameter tuning and regularization techniques.

Overall, the ability of XGBoost to handle complex relationships and its high predictive accuracy make it a suitable modelling technique for the data science salary prediction dataset.

ENSEMBLE MODEL OF BEST PERFORMING MODELS

The resulting model will be an ensemble of the most successful iterations of the top three models, with the objective of enhancing the accuracy of the final data science salary prediction model. By leveraging the strengths of each model, the ensemble approach is expected to mitigate the limitations of individual models and yield a more robust and accurate model. The ensemble technique has been widely recognized as an effective method for improving prediction accuracy and generalization performance in various machine learning applications.

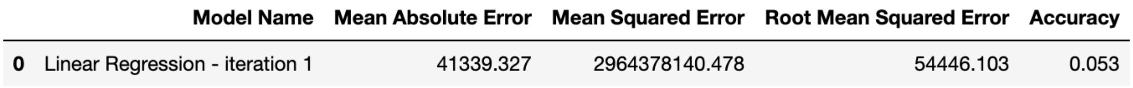
TRAIN SOLUTION

This section encompasses the training and iterative process of several modelling techniques for the data science salary prediction dataset. Due to the limited size of the dataset, a test-train split of 15-85 was deemed appropriate. Our primary objective is to scrutinize and enhance the performance of the top-performing models with the ultimate aim of devising a reliable and precise final model. To achieve this, we will employ several techniques, including hyperparameter tuning and feature selection, to optimize the performance of each model. The final selection of models will undergo meticulous evaluation, with emphasis on accuracy and generalization performance. The goal is to produce a comprehensive and robust ensemble model by integrating the best-performing models to maximize performance.

LINEAR REGRESSION

The first iteration of the linear regression model exhibited an accuracy score of only 0.053, which suggests that the model's performance is inadequate for the given data. The mean absolute error (MAE) of 41339.33 implies that the model's predictions were off by an average of $41,339. Similarly, the mean squared error (MSE) of 2964378140.48 indicates that the predictions were considerably dispersed and imprecise. The root mean squared error (RMSE) of 54446.10 further confirms that the model's predictions lacked accuracy.

Further improving the model's performance may involve tuning hyperparameters, such as the learning rate or regularization strength. However, considering the time constraints and the notably inadequate initial performance, it is prudent to entirely shift our focus to a different modelling technique. Figure 17 shows key metrics of the model's perf



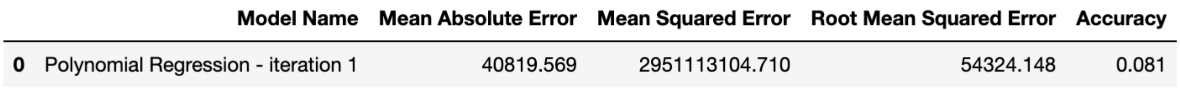
	Model Name	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Accuracy
0	Linear Regression - iteration 1	41339.327	2964378140.478	54446.103	0.053

Figure 17 - Metrics of Linear Regression Model

POLYNOMIAL REGRESSION

The first iteration of the polynomial regression model resulted in a slightly better accuracy score of 0.081, although it is still quite low for the given data. The mean absolute error (MAE) of 40819.569 suggests that the model's predictions were off by an average of $40,819. Similarly, the mean squared error (MSE) of 2951113104.710 indicates that the predictions were still dispersed and imprecise, albeit slightly better than the linear regression model. The root mean squared error (RMSE) of 54324.148 confirms that the model's predictions still lacked accuracy.

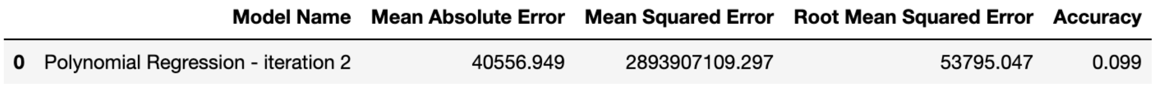
To improve the performance of the polynomial regression model, we may need to experiment with different degrees of the polynomial features this iteration is a 2nd order polynomial, as well as try different hyperparameters. It is possible that increasing the degree of the polynomial features may lead to better results. However, considering the time constraints and the relatively low accuracy score of the model, it may be worthwhile to explore other modelling techniques as well. Figure 18 shows key metrics of the models' performance.



	Model Name	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Accuracy
0	Polynomial Regression - iteration 1	40819.569	2951113104.710	54324.148	0.081

Figure 18 - Metrics of 2nd Order Polynomial Regression Model

Figure 19 shows the second iteration of the polynomial regression model improved significantly with an accuracy score of 0.099, indicating that the model's performance has substantially improved. The mean absolute error (MAE) decreased to 40556.949, which suggests that the model's predictions were off by an average of $40,556. Additionally, the mean squared error (MSE) decreased to 2893907109.297, indicating that the predictions were more precise and accurate. The root mean squared error (RMSE) decreased to 53795.047, further demonstrating the improvement in model performance. These results suggest that the 3rd order polynomial regression model may be a better fit for the given data than the linear regression model. Further analysis and experimentation may be necessary to optimize the model's performance, but the current results suggest that the polynomial regression model is a promising candidate for the task at hand.

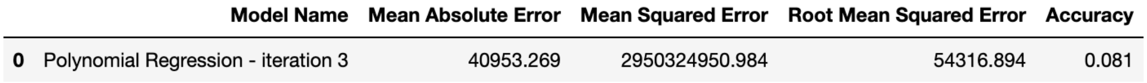


	Model Name	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Accuracy
0	Polynomial Regression - iteration 2	40556.949	2893907109.297	53795.047	0.099

Figure 19 - Metrics of 3rd Order Polynomial Regression Model

The third iteration of the polynomial regression model, a fourth-order polynomial, showed a slight decrease in performance compared to the previous iteration. The mean absolute error (MAE) of 40,953.27 suggests that the model's predictions were off by an average of $40,953. The mean squared error (MSE) of 2,950,324,950.98 indicates that the predictions were quite dispersed and imprecise. The root mean squared error (RMSE) of 54,316.89 confirms that the model's predictions lacked accuracy.

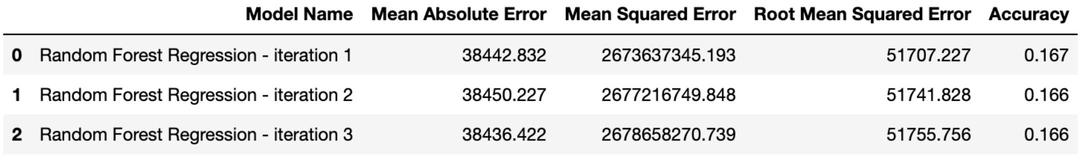
The decrease in performance may be attributed to overfitting, as the model's complexity increases with the addition of more polynomial features.



	Model Name	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Accuracy
0	Polynomial Regression - iteration 3	40953.269	2950324950.984	54316.894	0.081

Figure 20 - Metrics of 4th Order Polynomial Regression Model

RANDOM FOREST



	Model Name	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Accuracy
0	Random Forest Regression - iteration 1	38442.832	2673637345.193	51707.227	0.167
1	Random Forest Regression - iteration 2	38450.227	2677216749.848	51741.828	0.166
2	Random Forest Regression - iteration 3	38436.422	2678658270.739	51755.756	0.166

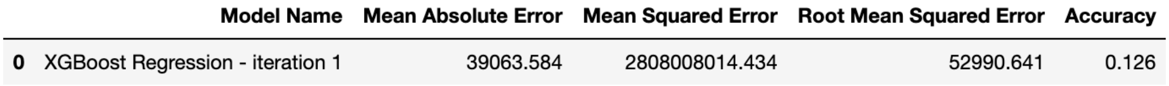
Figure 21 - Random Forest Modelling Metrics

The first iteration of the Random Forest Regression model demonstrated promising performance with an accuracy score of 0.167. The mean absolute error (MAE) of 38442.832 indicates that, on average, the model's predictions were off by $38,442. The mean squared error (MSE) 2673637345.193 suggests that the predictions were relatively dispersed. Additionally, the root mean squared error (RMSE) of 51707.227 highlights the model's ability to provide predictions with reasonable accuracy.

Random Forest Regression is an ensemble learning method that combines multiple decision trees to create a more robust model. In this case, the model utilized 100 estimators or decision trees. Increasing the number of estimators can potentially enhance the model's performance; however, it's essential to strike a balance as adding too many trees might lead to overfitting.

The second and third iterations of the random forest modelling employed an increased number of estimators at 200 and 800, respectively. Despite the greater complexity of these models, they exhibited a slight decline in accuracy and produced a similar mean absolute error (MAE) as compared to the initial iteration. These observations are illustrated in Figure 21, where the metrics for the first and subsequent iterations are presented.

GRADIENT BOOSTING (XGBOOST)

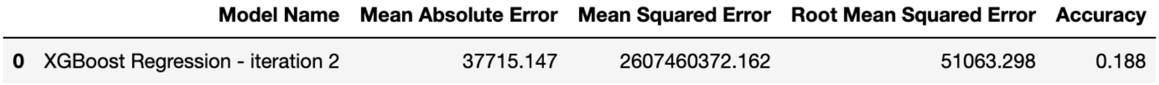


	Model Name	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Accuracy
0	XGBoost Regression - iteration 1	39063.584	2808008014.434	52990.641	0.126

Figure 22 - XGBoost Regression Model Iteration 1 Metrics

The XGBoost regression model with 100 estimators and a random state of 10 in its first iteration produced a mean absolute error of 39063.584, a mean squared error of 2808008014.434, a root mean squared error of 52990.641, and an accuracy score of 0.126. These results suggest that the model is not performing optimally, as the mean absolute error and root mean squared error are quite high, indicating that there is significant deviation between the predicted and actual values. However, the accuracy score is higher than that of the polynomial regression models, indicating that the XGBoost model is better at fitting the data.

To improve the performance of the XGBoost model, the number of estimators can be increased or the hyperparameters tuned. Increasing the number of estimators may improve the model's ability to capture the complex relationships within the data, but it may also lead to overfitting if the model becomes too complex. Hyperparameter tuning, on the other hand, involves adjusting the values of different model parameters to optimize its performance, such as the learning rate and maximum depth.



	Model Name	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Accuracy
0	XGBoost Regression - iteration 2	37715.147	2607460372.162	51063.298	0.188

Figure 23 - XGBoost Regression Model Iteration 2 Metrics

In the second iteration of XGBoost regression, a systematic hyperparameter search was conducted using GridSearchCV, a technique that explores various combinations of hyperparameters. The hyperparameters under consideration were the number of estimators, which determines the number of boosting rounds, the maximum depth of the individual decision trees within the ensemble, and the learning rate, which controls the contribution of each tree to the overall model. By exhaustively evaluating multiple combinations, the optimal hyperparameters were identified.

The refined XGBoost model derived from this process exhibited a modest reduction in the mean absolute error, mean squared error, and root mean squared error compared to the initial iteration. This decrease suggests an enhanced predictive performance, as the model's estimations were closer to the actual values. Furthermore, the accuracy of the model experienced a notable improvement, rising from 0.126 to 0.188. This increase in accuracy indicates an enhanced ability of the model to capture the underlying patterns and make more precise predictions of the target variable.

ENSEMBLE MODEL OF BEST PERFORMERS

In an attempt to further improve the accuracy and performance of the predictive model, an ensemble approach was employed. The top-performing models, as determined by their respective evaluation metrics, were selected to form the ensemble. Specifically, the ensemble model comprised the best iterations of XGBoost, Random Forest, and Polynomial Regression. These iterations were XGBoost - Iteration 2, Random Forest - Iteration 1, and Polynomial Regression - Iteration 2, respectively.

According to the results, the ensemble model achieved an MAE of 37894.578, an MSE of 2612352431.954, an RMSE of 51111.177, and an accuracy score of 0.187. The RMSE suggests that the predicted values have an average deviation of 51111.177, which is considerably high. Moreover, the accuracy score indicates that the model can explain only 18.7% of the variance in the target variable.

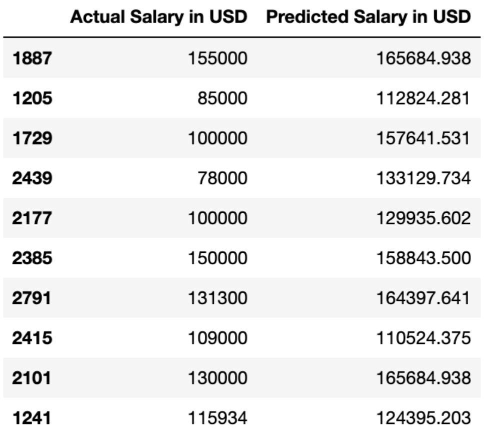
Regrettably, the bespoke ensemble model yielded inferior results compared to the XGBoost gradient boosting ensemble model. Enhancing the performance of the ensemble model would necessitate further investment of time and resources towards refining the individual models and conducting feature engineering. However, due to constraints on time, such measures could not be taken in the present study.

RESULTS & ANALYSIS

The XGBoost model from the second iteration was chosen to be subjected to further exploration and analysis. This was decided based on a thorough evaluation of the model's performance using appropriate metrics. Data visualization techniques will be utilized to scrutinize the results of the model and identify areas where it can be improved in future studies.

It's important to note that the task of predicting salary, which is a continuous form of data, is a challenging one. As such, it's difficult to achieve high accuracy in the predictions. However, the model was able to achieve a mean absolute error of $37715, indicating that the predicted salaries were typically within this range from the actual salary.

VISUALISATION & ANALYSIS OF MODEL PERFORMANCE



	Actual Salary in USD	Predicted Salary in USD
1887	155000	165684.938
1205	85000	112824.281
1729	100000	157641.531
2439	78000	133129.734
2177	100000	129935.602
2385	150000	158843.500
2791	131300	164397.641
2415	109000	110524.375
2101	130000	165684.938
1241	115934	124395.203

Figure 24 - Table showing 10 Actual Salaries against Predictions of Final Model

The analysis of the table displaying the actual and predicted salaries from the XGBoost model reveals varying levels of discrepancy. The predicted salaries differ from the actual salaries in different magnitudes across the rows.

Specifically, the predicted salaries are slightly higher than the actual salaries in the first and sixth rows, significantly higher in the second and third rows, considerably higher in the fourth row, and lower in the fifth row. The predicted salary in the seventh row is higher than the actual salary, while in the eighth row, it closely aligns with the actual salary. However, in the ninth and tenth rows, the predicted salaries are significantly higher than the actual salaries.

This analysis indicates that the XGBoost model may not accurately capture the underlying patterns and factors that influence salaries. To improve the accuracy of salary predictions, further analysis and refinement of the model are necessary. Potential approaches include hyperparameter tuning, feature engineering, and exploring alternative modelling techniques. These efforts aim to enhance the model's ability to capture the complexities and nuances of salary prediction.

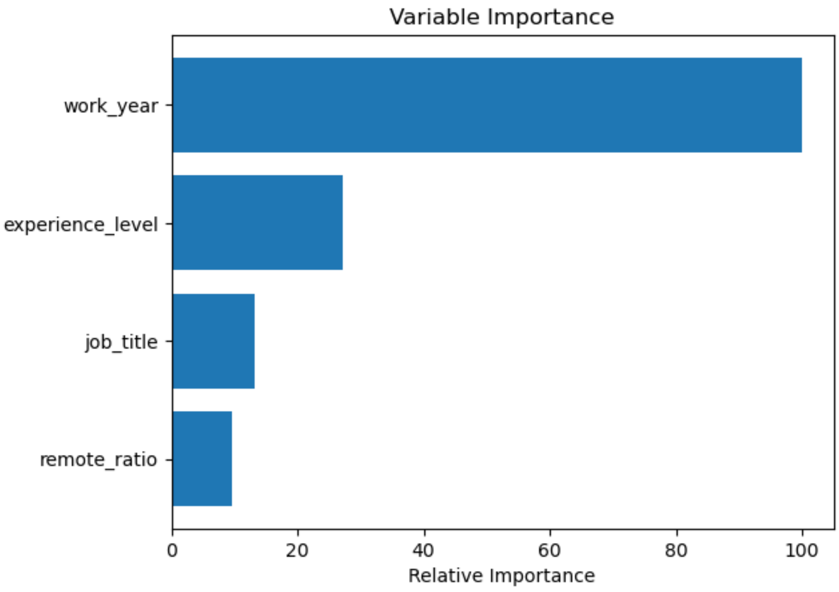


Figure 25 - Feature Importance Plot of Final Model

According to Figure 25, work year is identified as a highly important feature in the model. However, it is important to note that Figure 14 indicates only a moderate correlation between work year and salary. While the initial analysis acknowledged this correlation, it was also noted that there may be other complex factors influencing the relationship between salaries and work year such as the Covid-19 pandemic. Consequently, it is recommended that future iterations of the model should place less emphasis on work year and instead focus more on work experience, which exhibits a clear and discernible correlation with salary, as demonstrated in Figure 15. Such a refinement may help to improve the accuracy of the model and provide more nuanced insights into the relationship between these key features and the outcome variable.

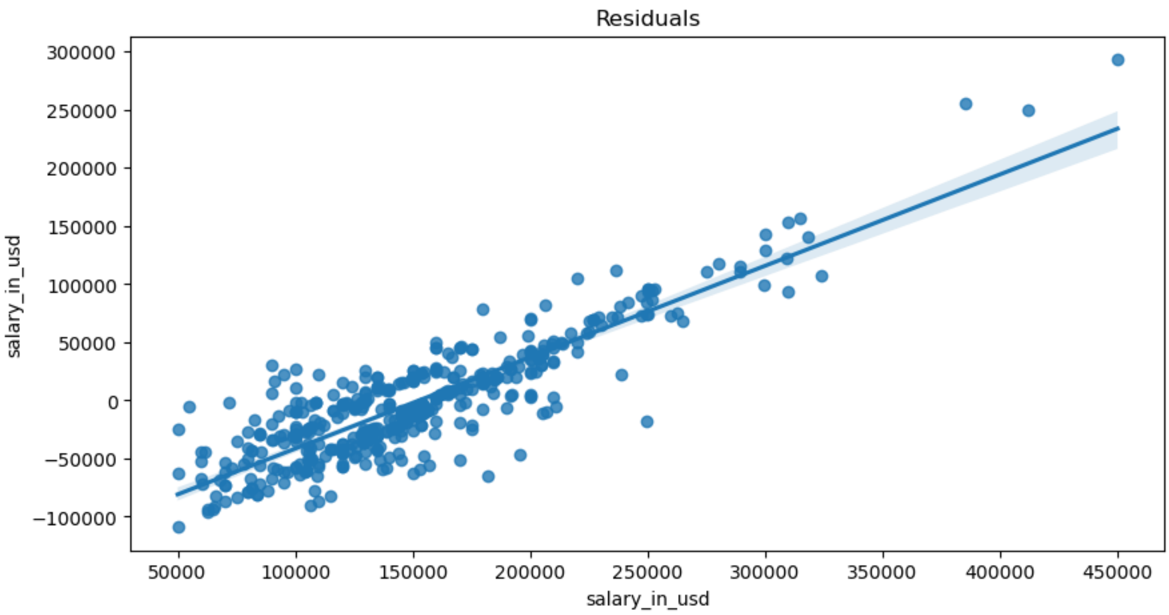


Figure 26 - Residuals Plot of Final Model

A residual plot serves as an essential graphical instrument to evaluate the performance of a model by depicting the deviations between observed and predicted values [8]. In such a plot, the x-axis displays the predicted values, while the y-axis indicates the residuals. Figure 26 reveals that the model tends to over forecast lower salaries and under forecast higher salaries, which significantly deviate from the average. However, the model exhibits higher accuracy for salaries ranging from 135,000 to 175,000. The plot suggests that the model fails to adequately capture the underlying patterns and trends in the data, especially for higher salaries. This graphical assessment is critical to understanding potential weaknesses and limitations of the model, guiding further improvements and modifications [9].

IMPLICATIONS OF SOLUTION AND CRITICAL ANALYSIS

The XGBoost model from the second iteration has demonstrated its potential in predicting salaries, although with varying levels of accuracy. The results obtained from the model and the visualizations, such as the table of actual and predicted salaries, feature importance plot, and residual plot, have provided valuable insights into the model's performance and areas for improvement. This section will critically analyse the implications of the solution and suggest future directions to enhance the model's performance.

Model Performance and Business Context

The model's mean absolute error of $37,715 indicates a certain level of accuracy, but it is essential to consider the business context when interpreting this result. Depending on the industry or application, this level of error may or may not be acceptable. In the technology industry there is a high level of variability in salaries and as such, a higher error might be tolerable to get a pall bark figure of what salary expectations should be. On the other hand, organizations seeking more precise salary predictions for budgeting or talent acquisition purposes might require a more accurate model.

Feature Selection and Engineering

The feature importance plot highlights the significance of work year in predicting salaries. However, as noted earlier, work year only exhibits a moderate correlation with salary. Future model iterations should explore alternative features, such as work experience or other features available in the original dataset such as company size, to improve the model's accuracy. Additionally, incorporating feature engineering techniques, such as creating interaction terms or using non-linear transformations, could help capture more complex relationships between variables.

Model Robustness and Generalizability

The residual plot demonstrates that the model performs better for certain salary ranges and struggles to capture the underlying patterns for higher salaries. This finding suggests that the model may not be robust and generalizable across different salary levels. Further investigation and refinement of the model are necessary to improve its performance and ensure its applicability to a broader range of salary predictions.

Hyperparameter Tuning and Alternative Modelling Techniques

To enhance the model's performance, future studies should explore hyperparameter tuning to optimize the XGBoost model. This process involves adjusting various model parameters to achieve the best possible performance on the given dataset. Additionally, further researchers should consider experimenting with alternative modelling techniques, such as deep learning, to identify the most suitable approach for predicting salaries in the given context.

In conclusion, the XGBoost model from the second iteration has demonstrated its potential for predicting salaries, although improvements are necessary to achieve higher accuracy and better generalizability. By addressing the identified limitations and incorporating the suggested improvements, future iterations of the model could provide more reliable and accurate salary predictions, offering valuable insights for organizations and individuals alike.

ML CHALLENGES AND LIMITATIONS

The prediction data set and modelling in this study encountered several challenges and limitations, which are typical in machine learning projects. These challenges and limitations impact the performance of the model and its applicability in real-world scenarios. This section will discuss some of the challenges and limitations that were encountered during this study.

DATA AVAILABILITY

The dataset used for the prediction of salaries initially comprised only 3755 rows, which underwent a rigorous data cleaning process that resulted in the removal of nearly 800 rows. This pre-processing step aimed to enhance the accuracy of the models developed. Nonetheless, the reduction in sample size due to data cleaning has reduced the overall representativeness of the data set. It is worth noting that with a larger dataset, there could have been more robust opportunities to split the dataset into training, testing, and validation sets, leading to a more accurate model. However, due to the limited size of the dataset, it was not feasible to set aside a considerable amount of data for validation. The paucity of features in the dataset also constrained the ability to remove weakly correlating features, as doing so could have had a profound impact on the models' performance. Despite the difficulties there are several possible mitigations that could be implemented in future studies to address these limitations such as, increasing the size of the dataset could potentially provide a more representative sample of the population, leading to more accurate models. This could be achieved by collecting data from multiple sources or by leveraging data augmentation techniques.

TIMING & RESOURCE AVAILABILITY

The temporal constraints associated with this project imposed certain limitations on the scope of model exploration, thereby precluding the ability to investigate more intricate models. This constraint was compounded by the fact that all model computations were conducted on a laptop, which is not an optimal computational resource for machine learning purposes. Consequently, this may have curtailed the potential depth and complexity of the model search and may have restricted the overall model performance in terms of accuracy and generalizability. To mitigate these constraints in future studies, researchers may consider utilizing cloud computing platforms that offer more powerful computing resources than a typical laptop. This approach would enable the exploration of more intricate models within a shorter timeframe. Additionally, researchers could consider automating the model exploration process by using machine learning pipelines, which would allow for efficient iteration over different models and parameters, thus enabling more comprehensive analysis within the available timeframe. Furthermore, collaborating with data scientists or experts in the field could facilitate the identification of optimal modelling techniques and ensure that the modelling process is conducted more efficiently. In conclusion, while this study was limited by time and resource constraints, the models developed can serve as a useful foundation for future research in salary prediction for the technology recruiting industry, and the suggested mitigations can potentially improve the efficacy and efficiency of the modelling process in future studies.

MODEL INTERPRETABILITY

The interpretability of a machine learning model is essential for understanding how the model makes predictions and identifying any potential biases in the model. However, certain machine learning algorithms, such as deep learning, can be difficult to interpret due to their complexity. In the present study, interpretability was achieved through the use of the XGBoost models feature importance plot and residual plot. While these techniques provided some insight into the model's performance, they may not provide a comprehensive understanding of the model's decision-making process. To address this limitation, future research could explore additional interpretability techniques, such as SHAP values or LIME, to gain a more comprehensive understanding of the model's underlying patterns and decision-making process. Incorporating these techniques may lead to improved accuracy and generalizability of the model and provide more nuanced insights into the relationship between the predictor variables and the outcome variable.

GENERALISABILITY

The ability of a model to generalize its performance beyond the original dataset is a vital aspect of its utility and practicality in real-world applications. In the present study, the model's performance was primarily evaluated on the available dataset. However, the generalizability of the model in predicting salaries for new, unseen data remains uncertain, particularly when the data significantly deviates from the original dataset. Therefore, it is essential to assess the model's generalizability across different contexts and data domains. Future research may entail examining the performance of the model on external datasets, cross-validation techniques, and statistical testing to evaluate the model's generalizability thoroughly. Moreover, the application of transfer learning or domain adaptation techniques may prove useful in enhancing the model's generalizability by leveraging existing knowledge from related domains. By addressing the generalizability limitations, the model's practical utility and applicability in diverse settings can be significantly enhanced.

ETHICAL CONSIDERATIONS

The development and deployment of the salary prediction model in this study warrant ethical considerations to mitigate the potential perpetuation of existing inequalities in the Data Science industry. The utilization of biased data or algorithms could exacerbate these disparities. Hence, ensuring that the dataset used for model training is representative and unbiased is paramount. Additionally, the model's deployment must account for fairness, accountability, and transparency. The model's transparency and interpretability are critical in enabling stakeholders to comprehend the underlying factors driving salary predictions and detecting potential biases in the model. Furthermore, it is crucial to ensure that the model's outputs are deployed in an ethical and unbiased manner that does not contribute to perpetuating social or economic disparities. However, it is worth noting that the absence of sensitive factors such as gender, race, and age in the dataset may limit the model's ability to account for and thus mitigate potential biases and discriminatory practices that may affect salary predictions. Thus, future iterations of the model should include such factors to increase its fairness and ethicality.

CONCLUSION

In conclusion, this paper presented an end-to-end machine learning project for predicting salaries in the data science industry using the Data Science Salary dataset. The project aimed to identify the factors influencing data science salaries and offered practical insights and techniques that can be applied to similar datasets in the future. The XGBoost model from the second iteration was found to be the best performing model, with a mean absolute error of $37,715, indicating that the predicted salaries were typically within this range from the actual salary. However, the model's performance was constrained by the limited size of the dataset and the lack of available features, which reduced its overall representativeness. Therefore, further exploration and refinement of the model are necessary to improve its generalizability and accuracy. Despite these limitations, our project highlights the potential of machine learning techniques in predicting salaries in the data science industry and the importance of end-to-end projects in providing a comprehensive understanding of the machine learning pipeline. Accurate salary predictions can aid in the hiring process, benefit both job seekers and employers, and help to promote a fair and equitable job market in the data science industry.

WORKS CITED

[1] Kaggle, “Data Science Salaries 2023,” 2023. [Online]. Available: https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023.

[2] I. Goodfellow, Y. Bengino and A. Courville, Deep Learning (Vol. 1), The MIT Press, 2016.

[3] I. Sarker, “Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective,” *Nature Public History Emergency Colection,* vol. 2, p. 377, 2021.

[4] M. Muller and S. Guido, Introduction to Machine Learning with Python: A Guide for Data Scientist, O'Reilly, 2016.

[5] R. Kumar and P. Yadav, “Salary Prediction Using Regression Techniques,” 2021.

[6] L. Breiman, “Random Forests,” *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[7] T. Chen and C. Guestrin, “XGBoost: A scalable Tree Boosting System,” in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[8] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, Springer, 2017.

[9] M. Kuhn and K. Johnson, Applied Predictive Modeling, Springer, 2013.

BIBLIOGRAPHY

Meadows, D., 2008. Thinking in Systems: A primer. Chelsea Green Publishing.

Geron, A., 2019. Hands-on Machine Learning with Scikit-Learn, Kera and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly

McKinney, W., 2017. Python for Data Analysis, 2e: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly