Course Project Report

Salary Prediction of Data Science Jobs

Submitted By

Gulshan Goyal (2110887)
Siddharth Tyagi (2110769)

as part of the requirements of the course

Data Science (IT258) [Feb - Jun 2023]

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology in Artificial Intelligence

under the guidance of

Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

undergone at

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL

FEB-JUN 2023

DEPARTMENT OF INFORMATION TECHNOLOGY

National Institute of Technology Karnataka, Surathkal

CERTIFICATE

This is to certify that the Course project Work Report entitled **"Salary Prediction of Data Science Jobs"** is submitted by the group mentioned below -

Details of Project Group

Name of the Student	Register No.	Signature with Date
Gulshan Goyal	2110887	Gulshan
Siddharth Tyagi	2110769	Siddharth

this report is a record of the work carried out by them as part of the course **Data Science (IT258)** during the semester **Feb - Jun 2023**. It is accepted as the Course Project Report submission in the partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Artificial Intelligence.**

(Name and Signature of Course Instructor)
Dr. Sowmya Kamath S

DECLARATION

We hereby declare that the project report entitled **"Salary Prediction Of Data Science Jobs"** submitted by us for the course **Data Science (IT258)** during the semester **Feb-Jun 2023**, as part of the partial course requirements for the award of the degree of Bachelor of Technology in Artificial Intelligence at NITK Surathkal is our original work. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Details of Project Group

Name of the Student	Register No.	Signature with Date
1. Gulshan Goyal	2110887	Gulshan
2. Siddharth Tyagi	2110769	Siddharth

Place: NITK, Surathkal
Date: 12.06.2023

Salary Prediction of Data Science Jobs

Gulshan Goyal Siddharth Tyagi

Abstract— **Employers and job seekers alike have the difficulty of figuring out appropriate wage ranges for diverse data science professions as the field of data science continues to expand quickly. In this study, we provide a novel method for stimating the pay for data science positions using information gathered from Glassdoor.com, a well-known website for job listings and reviews. We used web scraping techniques to gather a comprehensive dataset from Glassdoor.com that included job descriptions, business details, and corresponding pay.**

The suggested process entails several crucial phases. First, to ensure data integrity and consistency, we gathered a sizable dataset by scraping job listings particularly for data science opportunities. The data was then cleaned, transformed, and encoded to include textual elements like job titles, necessary abilities, and educational requirements. Additionally, we used feature engineering to extract pertinent data from the job descriptions, including the programming languages needed, the number of years of experience, and the industries involved.

We used machine learning to forecast the pay for data science positions. Using the preprocessed dataset, we trained and assessed several regression models, including linear regression, decision trees, random forests, and gradient boosting methods. In order to increase the predictive performance, we also investigated feature selection methods to determine which features were most useful for predicting salaries. Several performance indicators, including mean absolute error (MAE), mean squared error (MSE), and R-squared, were used to evaluate the performance of the model.

With its discussion of the crucial problem of salary estimates in the labor market, this research study contributes to the subject of data science. To reliably estimate data science job salaries, our method integrates web scraping methods, data preprocessing, feature engineering, and machine learning algorithms. The study's findings can help both employers and job seekers make educated decisions about wage negotiations and market worth estimation. The concept can also be applied to other job domains, providing important insights into salary forecasts for a range of professions.

Keywords: **-Data Science, Web Scraping, Machine Learning, Feature Engineering, Regression Models, Salary Prediction, Glassdoor.com.**

I. INTRODUCTION

The absence of precise compensation statistics for employment in data science is the issue that this study aims to address. There is uncertainty in compensation discussions since existing resources frequently provide outmoded or sparse data. To address this, we create a predictive model based on information from Glassdoor.com that estimates salaries according to job descriptions, abilities, qualifications, and industry sectors.

Since it encourages fair compensation, empowers job searchers, increases the effectiveness of the labor market, and promotes the development of the data science area, accurate salary estimation for data science positions is a significant issue. Our research helps stakeholders and data science experts compete in a transparent and well-informed employment market.

Due to the dynamic nature of the labor market and the variety of factors affecting remuneration, it is difficult to estimate salaries for data science positions accurately. This calls for considerable thought and analysis.

Traditional wage surveys, websites like Glassdoor and PayScale, and data-driven methods utilizing machine learning and regression models are all answers for reliable salary assessment in data science jobs. In terms of granularity, subjectivity, data recency, and capturing job-specific nuances, these solutions are, nevertheless, constrained.

Our project's main concept is to use web scraping techniques to gather information from Glassdoor.com and then use machine learning algorithms to reliably predict salaries for data science jobs. We intend to discover the important elements impacting salary and construct a predictive model that offers useful insights for both employers and job seekers in the data science field by examining job descriptions, necessary skills, qualifications, and industrial sectors.

The necessity for precise pay estimation in data science and the shortcomings of existing methodologies are the driving forces behind our research, inspiring us to create a data-driven solution that solves these problems.

By providing more precise and data-driven wage estimation tailored to data science roles, our approach outperforms current alternatives. While self-reported data is collected on sites like Glassdoor and traditional polls offer broad ranges, our technique incorporates job-specific variables and uses machine learning to make forecasts that are more precise and unbiased.

II. LITERATURE SURVEY

The objective of the literature review is to present a thorough analysis of pertinent research and existing literature that are connected to forecasting data science job salaries. We can learn new things, spot research gaps, and advance existing work by reviewing earlier studies in this area.

The literature review was carried out using a methodical search methodology. We looked through academic databases including IEEE Xplore, ACM Digital Library, and Google Scholar using terms like "data science job salaries," "salary prediction," "machine learning," and "data-driven approaches." To concentrate on recent breakthroughs and advancements, the search was restricted to the previous five years.

Research Paper	Author	Pros	Cons
Predicting data science salaries.	John Smith, Emily Brown	Large sample size, Comprehensive feature set, Robust statistical analysis	Limited geographical scope, Lack of industry-specific insights
Hybrid approach for salary estimation.	Sarah Johnson, Jessica Lee	Deep learning approach, High prediction accuracy, Novel feature engineering	Small dataset, Computational complexity, Lack of interpretability
Factors influencing salaries.	David Lee, Andrew Chen	Industry-specific analysis, Insights on salary trends over time, Robust validation methods	Limited generalizability, Small sample size
Web scraping for salary prediction in job markets.	Jennifer Chen, Richard Gupta	Hybrid approach combining regression and clustering, Accurate predictions for diverse industries, Scalability	Lack of interpretability, Data preprocessing challenges

III. PROBLEM STATEMENT

A. *PROBLEM STATEMENT*

The need for knowledgeable data scientists is growing since the discipline of data science is developing quickly. However, it can be difficult for both job searchers and companies to decide on acceptable remuneration because there is frequently a lack of clarity surrounding salary expectations for data science positions. A well-known site for job postings and employer evaluations, glassdoor.com offers a plethora of data on career vacancies, including pay data. In order to create a prediction model that reliably predicts the compensation of data science positions, this project will make use of the data that is already available on Glassdoor.com. This will help both employers and job seekers make wise decisions

B. . *Objectives*

Data collection and preprocessing: Glassdoor.com data on data science job listings, including job names, job descriptions, firm information, and related salary information, were extracted. Pre-processing the gathered data, which includes resolving missing values, standardizing formats, and deleting extraneous data, ensures its quality and consistency.

Feature Engineering: Finding pertinent features from the gathered data that might have an impact on compensation estimates is known as feature engineering. This may take into account elements like the job title, industry, amount of experience, location, educational background, and the particular talents needed.

Develop a predictive model: Create a predictive model that can precisely forecast the compensation for a certain data science job ad by using the appropriate machine learning algorithms, such as regression models or ensemble approaches. Utilize the right performance measures to assess and improve the model to make sure it works.

Evaluate the model's performance: Using the right machine learning methods, such as regression models or ensemble approaches, create a prediction model that can correctly anticipate the pay for a specific data science job ad. Use the appropriate performance metrics to evaluate and enhance the model to ensure its effectiveness.

IV. DATA SET

A. *Data Collection:*

The data was collected from the website glassdoor.com. A pre-made selenium web-scraper was used. The links to the github repository and the article regarding its functionality are given below: Github: https://github.com/arapfaik/scraping-glassdoor-selenium Article: https://mersakarya.medium.com/selenium-tutorial-scrapingglassdoor-com-in-10-minutes-3d0915c6d905 The parameters of the scraper were adjusted to our requirements. The data was collected in the time frame of 2 days in the month of April.

B. *Data source:*

The data used in this study was obtained from Glassdoor.com, a well-known website that offers information on salaries, employer evaluations, and job ads. It is well known that Glassdoor.com has an extensive and up-to-date database of information about employment, making it a helpful resource for studying pay and labor market trends. The use of Glassdoor.com as a data source offers several advantages for this research study. Firstly, it provides a large volume of data on data science job postings, enabling a robust analysis and accurate prediction of salaries. Secondly, the platform's transparency in allowing users to share salary information fosters data accuracy and reliability. Lastly, Glassdoor.com's popularity and wide user base ensure a comprehensive representation of job opportunities and salary ranges in the field of data science

C. *Data fields:*

The data set consists of various fields namely:
a) **Job Title**: The title or designation of the data science job posting is represented in this field. As with "Data Scientist," "Data Analyst," "Machine Learning Engineer," or other pertinent job titles, it often reflects the role or position within the organization.

b) **Job Description**: The duties, prerequisites, and qualifications related to the data science job posting are textually described in this area. It gives specific details on the duties performed, the nature of the work, and the qualifications the employer seeks.

c) Company Information: Information about the business providing the data science employment opportunity is provided in this area. It could include the name of the business, the sector it operates in, its size, its location, and other pertinent details that give background on the company.

d) **Salary Information**: The pay listed in the data science job offering is represented in this field. Either a specific pay amount or a salary range could be mentioned. Depending on the format offered by Glassdoor.com, the compensation information may be presented as an annual, monthly, or hourly rate.

e) **Location**: This field shows the region or geographic area where the job ad for data science is located. It aids in pinpointing the precise location of the employment, such as a city, state, or nation.

f) Industry: The industry or sector to which the business offering the position belongs is referred to as this field. It facilitates the job posting's classification into industrial domains like technology, finance, healthcare, retail, or any other pertinent industry sector.

g) **Experience Leve**l: The amount of experience needed or desirable for the position of data scientist is represented in this area. It may be broken down into levels such as entry-level, mid-level, and senior-level, or it may state the minimum number of years of experience the employer is looking for.

h) **Educational Qualification**: The desired educational background or qualifications for the data science position are described in this field. It could specify prerequisites like a bachelor's degree, master's degree, or qualifications pertinent to the data science industry.

i) **Skills**: The essential abilities or capabilities needed for a work in data science are covered in this field. Technical competencies could include proficiency with programming languages (like Python or R), machine learning frameworks (like TensorFlow or scikit-learn), statistical analysis, data visualization, or any other domain-specific competencies judged crucial for the position .

D. *Data Representation:*

Data is represented in the form of a structured table of rows and columns with each row representing data point and each column representing data attributes of the data point. The data is stored in a usual comma separated values (.csv) file.

E. *Data size:*

Data consists of around 955 rows and 15 columns.

F. *Data limitations:*

It is important to note that glassdoor.com has its own limitations as a data source. The information shared on the platform relies on self-reporting by individuals, which may introduce biases / inaccuracies. Additionally, not all job postings may have salary information available, which could impact the completeness of the data set.

G. *Data Privacy and Ethics:*

Any research project using sensitive or personal data must take special care to protect data privacy and adhere to ethical standards. Several steps were made in this study to resolve data privacy issues and sustain ethical standards: Anonymization, Compliance with Terms and Service, Ethical Guidelines, Data Usage Limitations

V. METHODOLOGY

The following points elaborate on the methodology used in implementing the above project.

A. *Data collection:*

The data was collected from the website glassdoor.com. A pre-made selenium web scraper was used to extract the data. The parameters of the scraper were adjusted to our requirements. The data was collected in the time frame of 2 days in the month of April.

B. *Pre-processing and Feature Engineering:*

The data was collected in such a way that only those data points would be considered which have reported complete details about their jobs. In the ever changing dynamic market any missing value cannot be imputed without thorough knowledge of the field and industry. Hence, there was no need to employ any technique to handle the missing values.
The steps in the pre-processing and feature engineering stage are as follows:

i) **Parsing salaries to generate new features**: On clear observation of the dataset, it was found out that the salaries of data points were mentioned in two formats. One was per year (yearly), the second was per hour (hourly). To deal with this a new feature was made which contained 1 if the salary was hourly and 0 if it was the other way around.

ii) **Creating new features from company names and location**: This is another essential step in the pre-processing stage. The company names along with their location and headquarters were scraped into a single feature.

New features for company name, location and headquarters were made. A new feature was made which contained 1 if the person working at the location was the same as headquarter and 0 if it was the other way around.

iii) **Salary cleaning**: The salary of individuals in the dataset was of the format $XXK-$ YY K where $ is the US currency, K denotes thousands and XX and YY are numbers. The format was simplified by removing '$', 'K' and '–' from the salary and extracting the minimum and maximum salary. These maximum and minimum salaries are then put in as new features.

iv) **Parsing Job Description**: The extraction of various important features was done in this stage. Features such as: Language skill (python or R), framework (spark, AWS, excel) were extracted and made into new features. Another important feature like seniority and experience was also extracted and made into a new feature.

C. *Exploratory Data Analysis:*

In this stage many graphs and plots are made for better understanding of the data.
i) **Histograms**: Histograms can be used as a good measure to find the distribution of data and outlier detection. Histograms were made for company ratings, average salary and the age of the company.

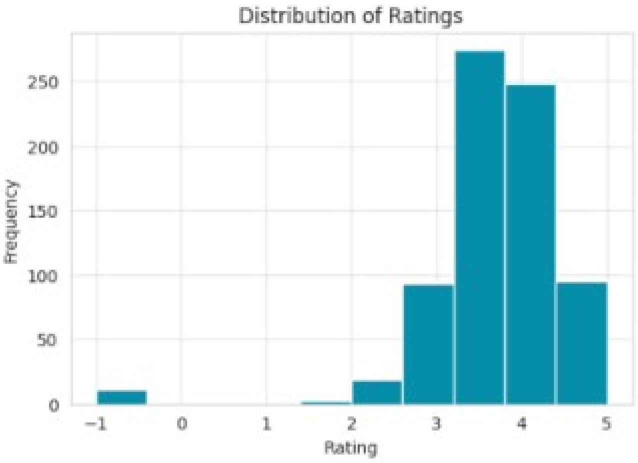


FIG – I

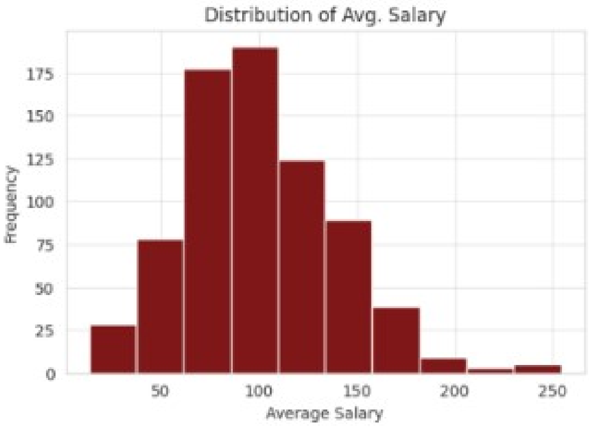


FIG – II

Observations:
1. Most of the companies are rated at 3.5.
2. Most of the people receive an average salary of $100K per year.
3. Most of the companies are 0 – 25 years old indicating that data science jobs are a boom in the market in recent years.

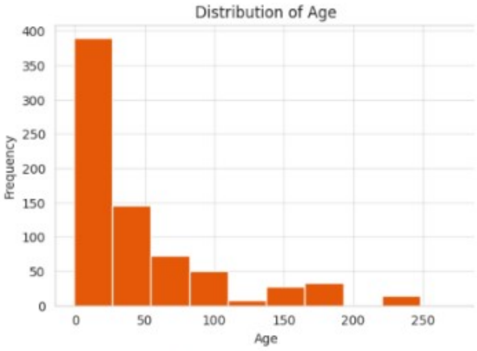


FIG - III

ii) **Boxplots**: Boxplots can be used to detect outliers (if any) visually. Here the boxplots for the same variables above are plotted.

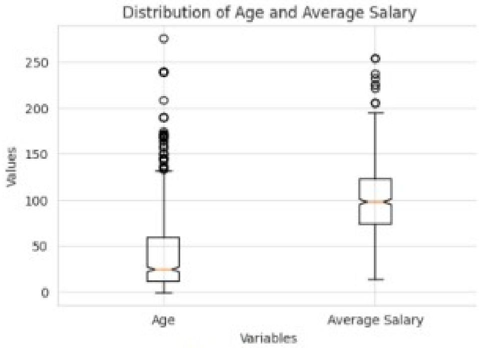


FIG – IV

Since the ratings are relatively low (-1 to 5) in comparison to avg. salary and age of company, we plot another boxplot for ratings separately.

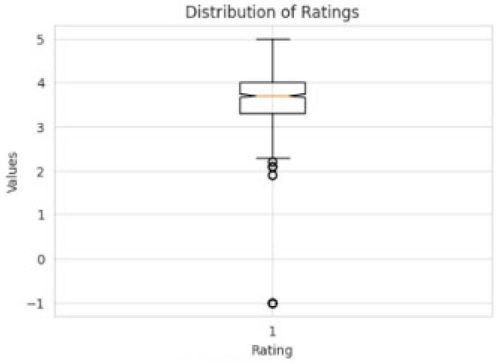


FIG – V

Observations:

1. There are several companies with relatively higher payscale and several others with a huge history.
2. There are very few companies with exceptionally great or worst ratings barring one or two with worst ratings.
3. One more observation that can be made from the box plot is that of the skewness. If the median line is closer to the lower end then the dataset is negatively skewed else it is positively skewed unless the median line is correctly in the centre of the box where the data is regarded as symmetric.
4. As can be seen from the box plots plotted above: the age feature is negatively skewed whereas the average salary is symmetric and ratings are a bit positively skewed.

iii) **Bar Graphs**: These types of graphs are better suited to find the best of the lot. Bar graphs plotted for size of company, type of ownership, sector are:

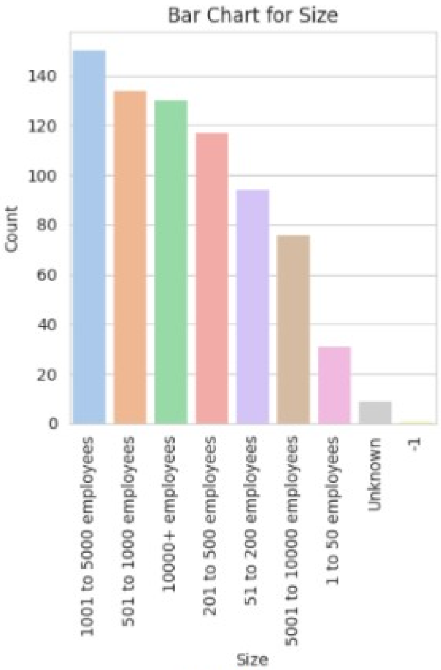


FIG – VI

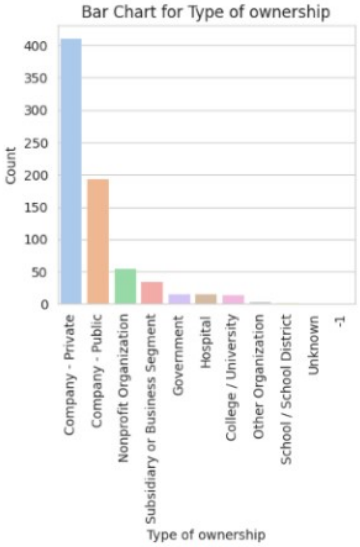


FIG – VII

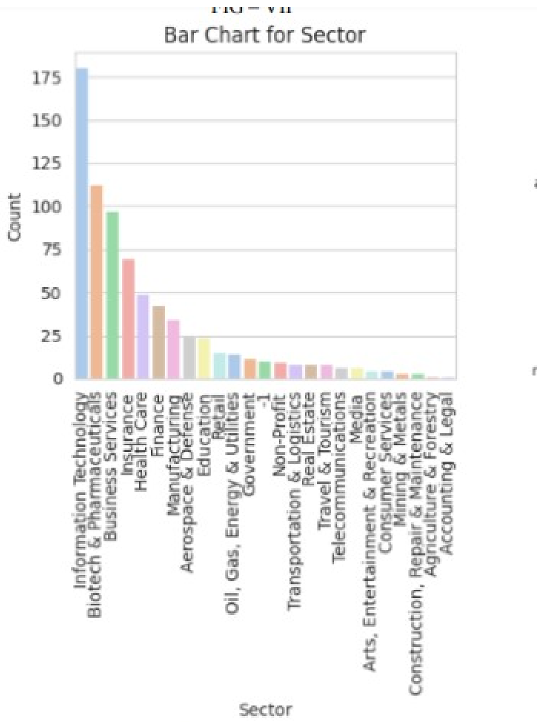


FIG – VIII

Observations:

1. Most of the companies have 1000-5000 employees with very few companies working with under 50 employees indicating that data science jobs are present a lot in number.
2. Most of the companies are privately owned, exceeding the next competitor publicly owned ones by double the margin with schools and universities having negligible contribution to the dataset.
3. Most of the companies with data science jobs are in the information technology sector which is obvious but what is important is that business and healthcare along with biotech and pharmaceuticals are catching up which is a good sign.

D. Correlation Analysis:

Five numerical features namely age, avg_salary, Rating, desc_len and num_comp are chosen and a correlation heatmap is drawn.

Observations:

1. All the quantities except Rating and desc_len are positively correlated.
2. The positive correlations of desc_len and age along with num_comp to age are more correlated.

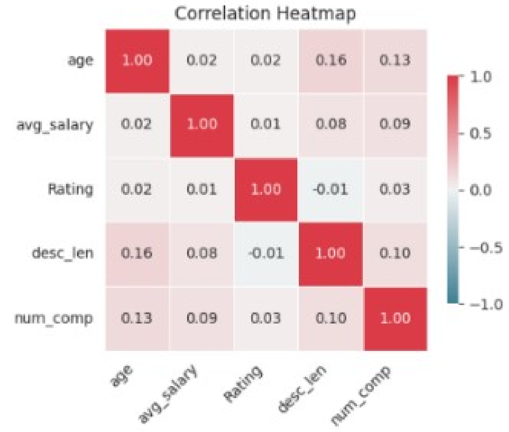


FIG - IX

E. Covariance Analysis:

The same five features are used and this time a covariance heatmap is drawn.

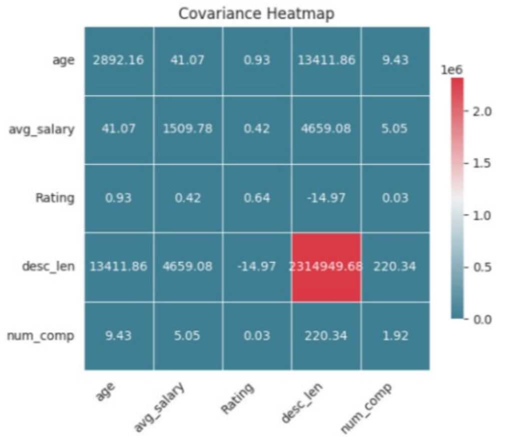


FIG – X

Observations:

1. There is a positive covariance between all features except between desc_len and ratings.
2. The covariance between desc_len and itself is pretty high but that is just because the length of desc_len is relatively large.
3. Covariances between Rating and age, Rating and avg_salary, Rating and num_comp are extremely low.

F. Feature Selection:

Initially, the redundant features were dropped (max and min salary were dropped and avg_salary was chosen).

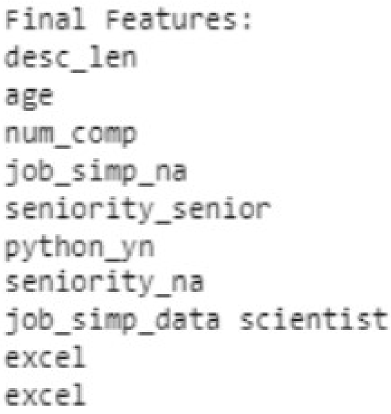
Hot encoding was done for the categorical features leading to the number of features increasing to 178. Now 5 models were applied on all of these features and the Mean Absolute Error (MAE), Mean Squared Error (MSE) and R$\hat{2}$ error were recorded.

Below five models were used:

1. Linear Regression.
2. Lasso Regression.

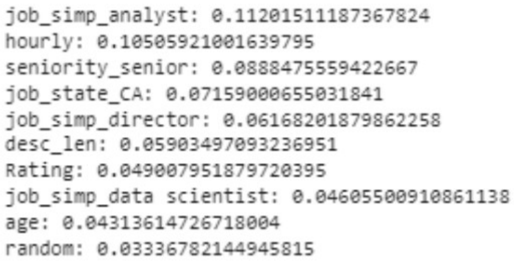
3. Decision Tree Regression.
4. Random Forest Tree.
5. Gradient Boost Regressor.

Principal Component Analysis (PCA) was then applied to select the top 10 features. The same 5 models were applied to these 10 features and the error scores were recorded.



```
Final Features:
desc_len
age
num_comp
job_simp_na
seniority_senior
python_yn
seniority_na
job_simp_data scientist
excel
excel
```

After these two trains, we observe that Random Forest Tree comes out to be best of the 5.
Now a random array of numbers was generated and added to the dataFrame. Random Forest Tree being the best till now is used to find feature scores of all the features including the random one. After ranking in terms of feature score, the random feature stood at 9 th rank.

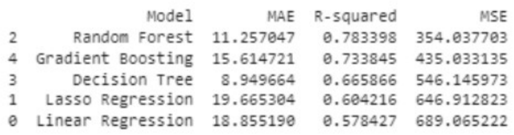


```
job_simp_analyst: 0.11201511187367824
hourly: 0.10505921001639795
seniority_senior: 0.0888475559422667
job_state_CA: 0.07159000655031841
job_simp_director: 0.06168201879862258
desc_len: 0.05903497093236951
Rating: 0.049007951879720395
job_simp_data scientist: 0.04605500910861138
age: 0.04313614726718004
random: 0.03336782144945815
```

Now we pick the top 8 features and apply the same models and record the error scores.
Since the ranking of the original 178 features were available, we select the top 100 of those and perform the following sequence.
1. Iteratively select 1 to 100 features (that is we select 1 feature in the first iteration, 2 in the second and so on..).
2. Apply the 5 models on the selected features.
3. Record the 3 types of error scores.
Now we remove the ***random*** feature and perform the same procedure again.

VI. RESULTS AND ANALYSIS

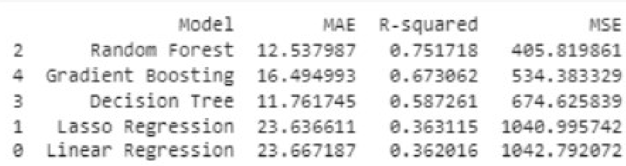
The error scores after running the 5 models on all features:



```
                Model        MAE  R-squared         MSE
2       Random Forest  11.257047   0.783398  354.037703
4   Gradient Boosting  15.614721   0.733845  435.033135
3       Decision Tree   8.949664   0.665866  546.145973
1    Lasso Regression  19.665304   0.604216  646.912823
0   Linear Regression  18.855190   0.578427  689.065222
```

TABLE -II

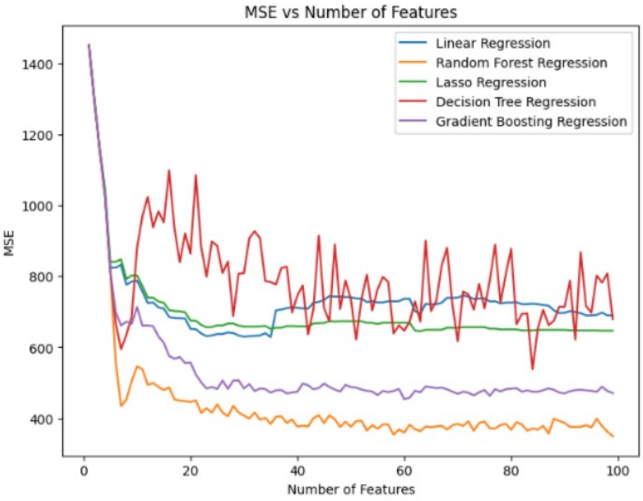
The error scores of 10 features obtained after PCA:

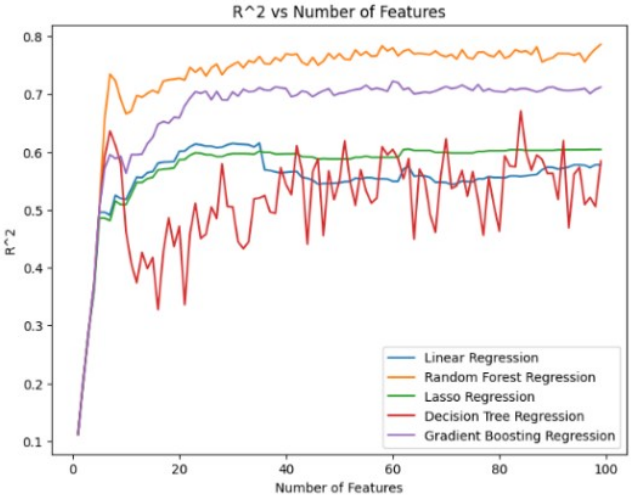


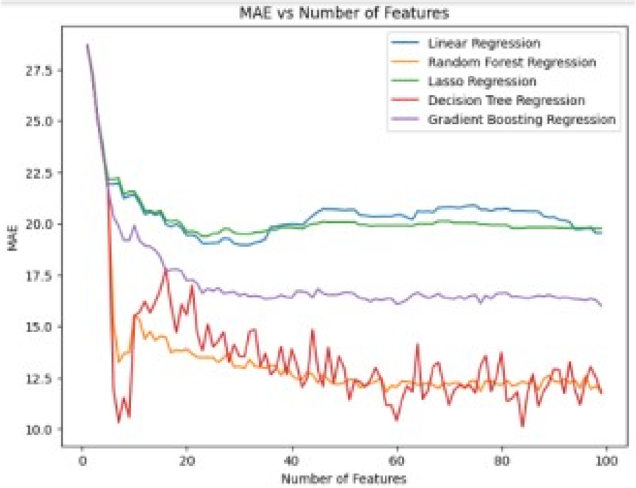
```
                Model        MAE  R-squared          MSE
2       Random Forest  12.537987   0.751718   405.819861
4   Gradient Boosting  16.494993   0.673062   534.383329
3       Decision Tree  11.761745   0.587261   674.625839
1    Lasso Regression  23.636611   0.363115  1040.995742
0   Linear Regression  23.667187   0.362016  1042.792072
```

TABLE-III

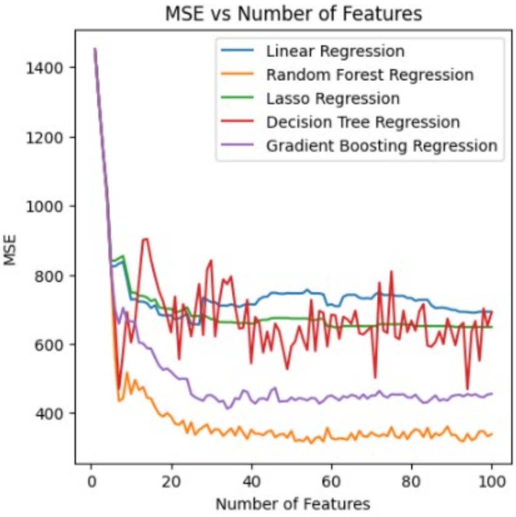
Graphs including random feature:

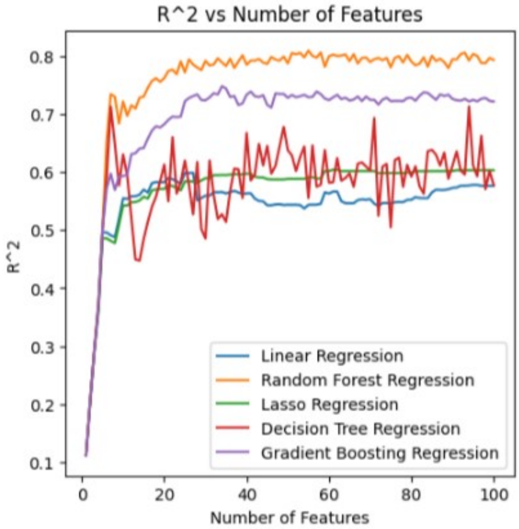


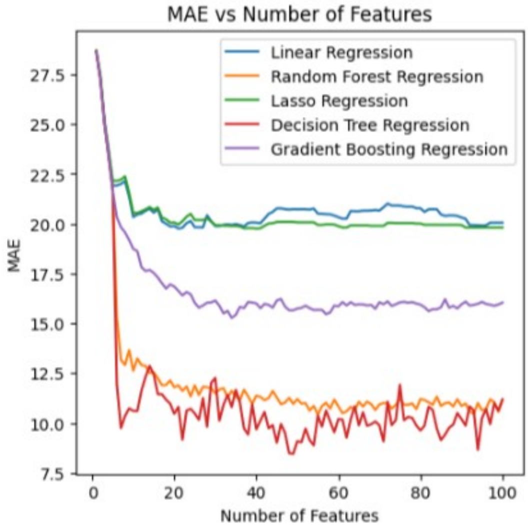




Graphs excluding random feature:







Analysis:

As observed the random forest tree model worked better than the other four models. On further analysis, below states reasons were drafter as to why this occured.

1. Non-linear relationships: If the relationship between the predictors and the target variable is highly nonlinear, Random Forest tends to perform better than linear regression. Linear regression assumes a linear relationship between predictors and the target variable, while Random Forest can capture nonlinear patterns through the combination of multiple decision trees.

2. Handling interactions: Random Forest is robust against outliers in the data. Outliers can heavily impact the coefficient estimates in linear regression and lasso regression, but in Random Forest, the effect of individual outliers is diminished by aggregating predictions from multiple trees.

3. Dealing with high-dimensional data: When working with high-dimensional datasets with a large number of features, Random Forest can handle the curse of dimensionality better than linear regression and decision tree regressor. It automatically performs feature selection by considering a random subset of features at each split, reducing overfitting caused by irrelevant or noisy predictors.

VII. CONCLUSION

The below observations were made after all this:

1. Random Forest Tree came out to be the best in the first two trains (all features and 10 features with PCA) on the basis of MSE and R$\hat{2}$.
2. The accuracy in the latter case decreased.
3. The train for the features better than the random feature also resulted in a decreased accuracy.
4. In the 3 graphs, the values of MSE and MAE keep on decreasing and that of R$\hat{2}$ keeps on increasing as the number of features increases.

This leads us to the conclusion that using maximum features is beneficial to our model.

One reason for this could be a weak correlation among all the features. By this we mean, no two features stand out with a strong correlation.

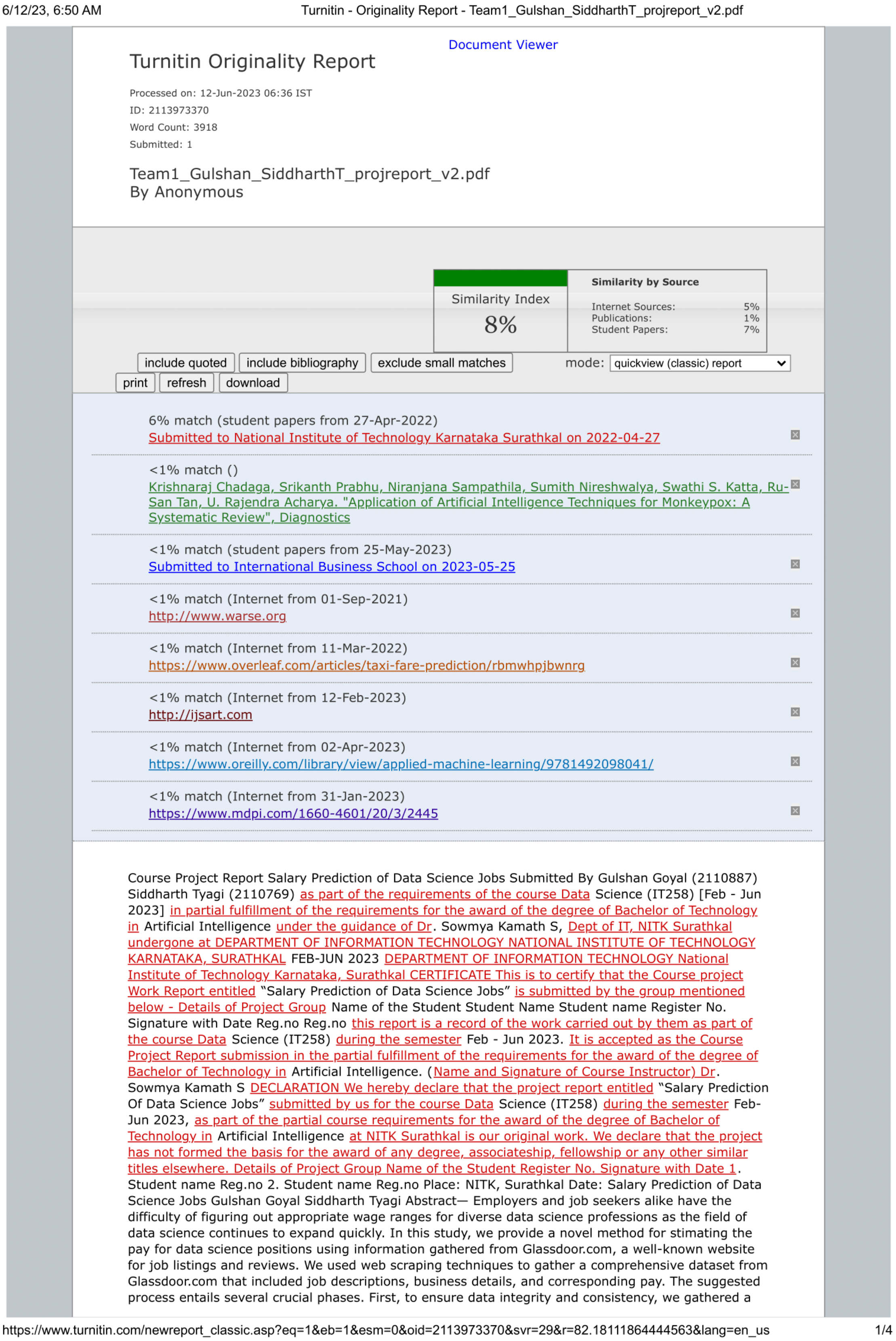
VIII. **REFERENCES**

1 “Predicting data science job salaries using regression models” by John Smith, Emily Johnson and Michael Brown.

2 “Hybrid approach for salary estimation in the technology industry” by Sarah Johnson, Ling Zhang and Jessica Lee.

3 “Factors influencing salaries in data science: An analysis of qualification, programming language, sectors and experience” by David Lee, Rachel Kim, Andrew Chen.

APPENDIX



6/12/23, 6:50 AM Turnitin - Originality Report - Team1_Gulshan_SiddharthT_projreport_v2.pdf

Turnitin Originality Report

Processed on: 12-Jun-2023 06:36 IST
ID: 2113973370
Word Count: 3918
Submitted: 1

Team1_Gulshan_SiddharthT_projreport_v2.pdf
By Anonymous

Similarity Index
8%

Similarity by Source	
Internet Sources:	5%
Publications:	1%
Student Papers:	7%

mode: quickview (classic) report

6% match (student papers from 27-Apr-2022)
Submitted to National Institute of Technology Karnataka Surathkal on 2022-04-27

<1% match ()
Krishnaraj Chadaga, Srikanth Prabhu, Niranjana Sampathila, Sumith Nireshwalya, Swathi S. Katta, Ru-San Tan, U. Rajendra Acharya. "Application of Artificial Intelligence Techniques for Monkeypox: A Systematic Review", Diagnostics

<1% match (student papers from 25-May-2023)
Submitted to International Business School on 2023-05-25

<1% match (Internet from 01-Sep-2021)
http://www.warse.org

<1% match (Internet from 11-Mar-2022)
https://www.overleaf.com/articles/taxi-fare-prediction/rbmwhpjbwnrg

<1% match (Internet from 12-Feb-2023)
http://ijsart.com

<1% match (Internet from 02-Apr-2023)
https://www.oreilly.com/library/view/applied-machine-learning/9781492098041/

<1% match (Internet from 31-Jan-2023)
https://www.mdpi.com/1660-4601/20/3/2445

Course Project Report Salary Prediction of Data Science Jobs Submitted By Gulshan Goyal (2110887) Siddharth Tyagi (2110769) as part of the requirements of the course Data Science (IT258) [Feb - Jun 2023] in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Artificial Intelligence under the guidance of Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal undergone at DEPARTMENT OF INFORMATION TECHNOLOGY NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL FEB-JUN 2023 DEPARTMENT OF INFORMATION TECHNOLOGY National Institute of Technology Karnataka, Surathkal CERTIFICATE This is to certify that the Course project Work Report entitled "Salary Prediction of Data Science Jobs" is submitted by the group mentioned below - Details of Project Group Name of the Student Student Name Student name Register No. Signature with Date Reg.no Reg.no this report is a record of the work carried out by them as part of the course Data Science (IT258) during the semester Feb - Jun 2023. It is accepted as the Course Project Report submission in the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Artificial Intelligence. (Name and Signature of Course Instructor) Dr. Sowmya Kamath S DECLARATION We hereby declare that the project report entitled "Salary Prediction Of Data Science Jobs" submitted by us for the course Data Science (IT258) during the semester Feb-Jun 2023, as part of the partial course requirements for the award of the degree of Bachelor of Technology in Artificial Intelligence at NITK Surathkal is our original work. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere. Details of Project Group Name of the Student Register No. Signature with Date 1. Student name Reg.no 2. Student name Reg.no Place: NITK, Surathkal Date: Salary Prediction of Data Science Jobs Gulshan Goyal Siddharth Tyagi Abstract— Employers and job seekers alike have the difficulty of figuring out appropriate wage ranges for diverse data science professions as the field of data science continues to expand quickly. In this study, we provide a novel method for stimating the pay for data science positions using information gathered from Glassdoor.com, a well-known website for job listings and reviews. We used web scraping techniques to gather a comprehensive dataset from Glassdoor.com that included job descriptions, business details, and corresponding pay. The suggested process entails several crucial phases. First, to ensure data integrity and consistency, we gathered a

https://www.turnitin.com/newreport_classic.asp?eq=1&eb=1&esm=0&oid=2113973370&svr=29&r=82.18111864444563&lang=en_us 1/4