MSBA SUMMER 2023

PREDICTING MOST LUCRATIVE DATA SCIENCE JOBS

Data Science Programming – Group Project

Group 19

The University of Texas at Austin

TEAMMATES

Jiarui Chang

Santhosh Kumar

Vi Tran

Meenal Gaba

AGENDA

- Background and Objectives
- Understanding the dataset and cleaning
- Exploratory data analysis
- Predictive Modeling
- Recommendations and conclusion

Background

- Data science is one of the hottest job profiles in the market right now and it is experiencing remarkable growth over the years
- In 2022, the data and analytics industry is worth ***USD ~250 billion and it is expected to grow to USD ~330 billion (↑35% increase)***
- The U.S. Bureau of Labor Statistics predicts there will be ***~11.6 million data science-related jobs in 2026***
- Irrespective of their educational backgrounds, people are shifting towards data science due to its relevance and attractive salary
- Understanding the variation in this market due to multiple factors like ***job profile, size of the company, location of company, experience, mode of work*** etc., will be key to understanding the earning potential

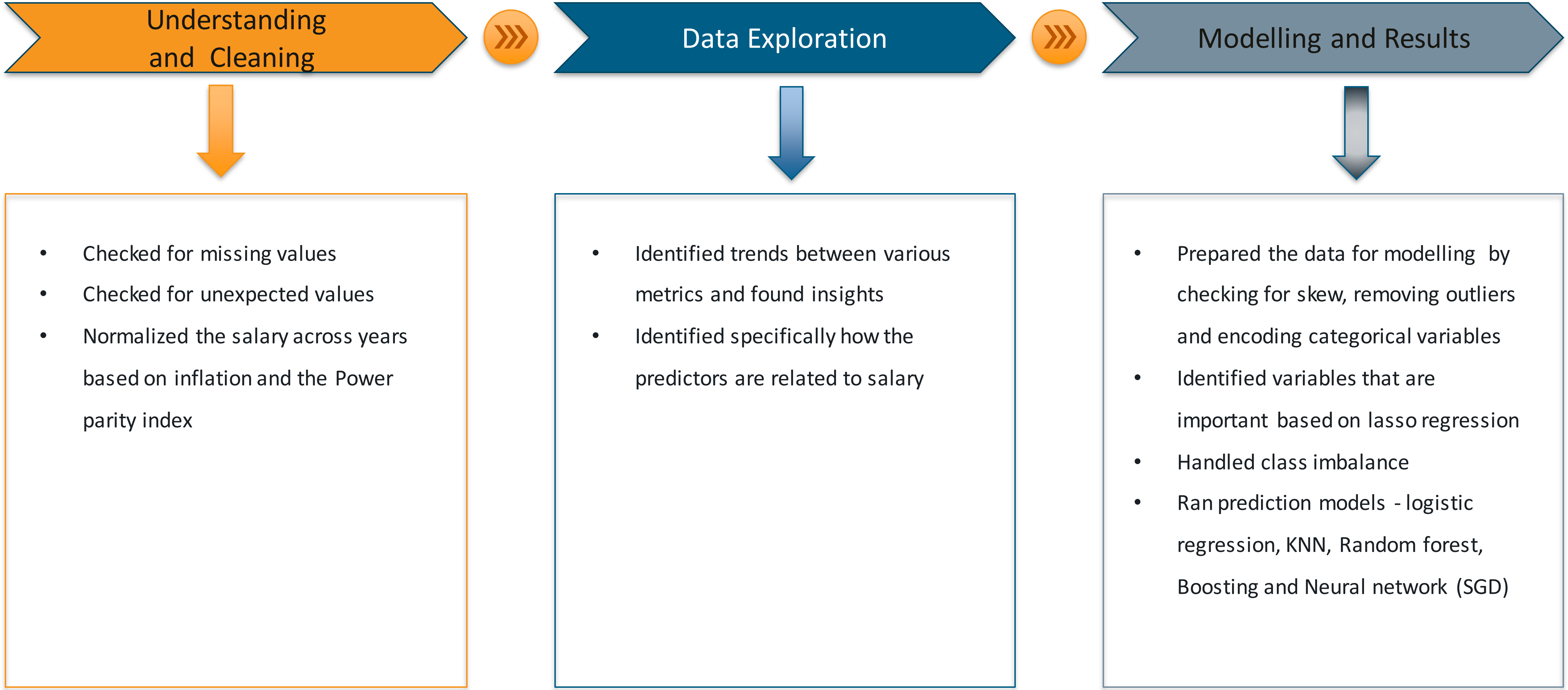
Objectives

- Understanding the lay of the land in terms of how different metrics influence each other in the field of data science jobs
- Figuring out the metrics that have an impact on the salary
- Predicting the most lucrative jobs within the data science domain - ***Who are the data scientists that earn more the 75% of people in this domain?***

AGENDA

- Background and Objectives
- Understanding the dataset and cleaning
- Exploratory data analysis
- Predictive Modeling
- Recommendations and conclusion

We have followed a three-step process to arrive at the solution



Understanding and Cleaning

- Checked for missing values
- Checked for unexpected values
- Normalized the salary across years based on inflation and the Power parity index

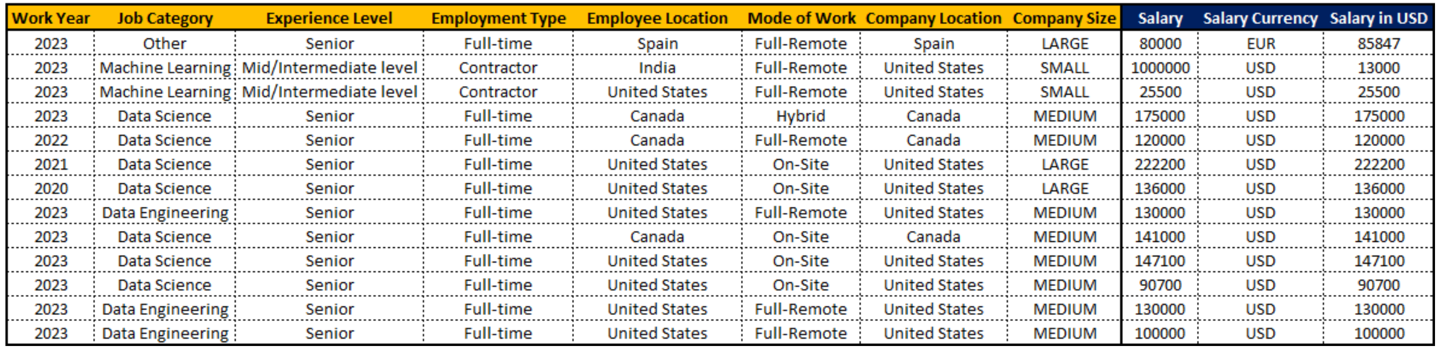
Data Exploration

- Identified trends between various metrics and found insights
- Identified specifically how the predictors are related to salary

Modelling and Results

- Prepared the data for modelling by checking for skew, removing outliers and encoding categorical variables
- Identified variables that are important based on lasso regression
- Handled class imbalance
- Ran prediction models - logistic regression, KNN, Random forest, Boosting and Neural network (SGD)

Data Snapshot



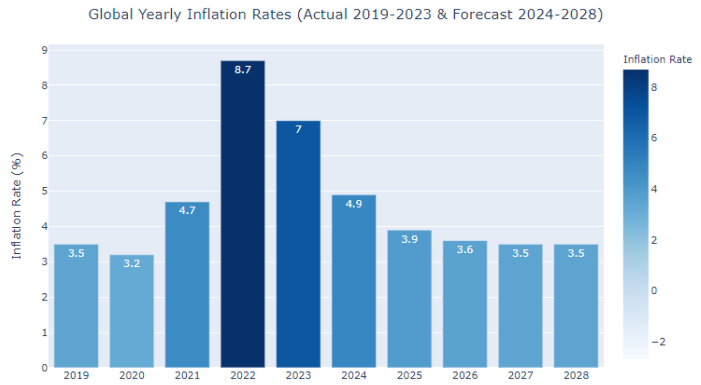
Work Year	Job Category	Experience Level	Employment Type	Employee Location	Mode of Work	Company Location	Company Size	Salary	Salary Currency	Salary in USD
2023	Other	Senior	Full-time	Spain	Full-Remote	Spain	LARGE	80000	EUR	85847
2023	Machine Learning	Mid/Intermediate level	Contractor	India	Full-Remote	United States	SMALL	1000000	USD	13000
2023	Machine Learning	Mid/Intermediate level	Contractor	United States	Full-Remote	United States	SMALL	25500	USD	25500
2023	Data Science	Senior	Full-time	Canada	Hybrid	Canada	MEDIUM	175000	USD	175000
2022	Data Science	Senior	Full-time	Canada	Full-Remote	Canada	MEDIUM	120000	USD	120000
2021	Data Science	Senior	Full-time	United States	On-Site	United States	LARGE	222200	USD	222200
2020	Data Science	Senior	Full-time	United States	On-Site	United States	LARGE	136000	USD	136000
2023	Data Engineering	Senior	Full-time	United States	Full-Remote	United States	MEDIUM	130000	USD	130000
2023	Data Science	Senior	Full-time	Canada	On-Site	Canada	MEDIUM	141000	USD	141000
2023	Data Science	Senior	Full-time	United States	On-Site	United States	MEDIUM	147100	USD	147100
2023	Data Science	Senior	Full-time	United States	On-Site	United States	MEDIUM	90700	USD	90700
2023	Data Engineering	Senior	Full-time	United States	Full-Remote	United States	MEDIUM	130000	USD	130000
2023	Data Engineering	Senior	Full-time	United States	Full-Remote	United States	MEDIUM	100000	USD	100000

Data Dictionary

Variable	Description	Variable	Description
Work Year	The year the salary was paid (2020 - 2023)	Employee Location	Employee's country of residence during the work year
Job Category	Data Science | ML Engineer | Data Architect | Data Engineer | Others	Company Location	The country of the employer's office
Experience Level	Senior | Mid | Entry	Mode of work	On-Site | Hybrid | Remote
Employment Type	Full time | Contractor | Part time | Freelance	Company Size	The median number of people that worked for the company during the year (Large | Medium | Small)
Salary | Salary Currency	The total gross salary amount paid and local currency	Salary in US	The salary in USD

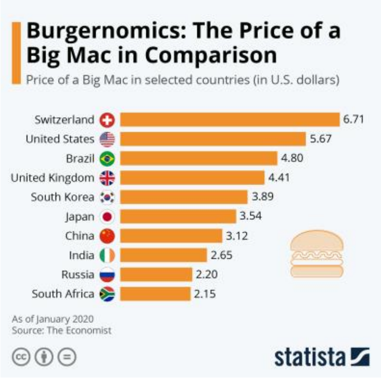
Normalized salaries based on inflation rates and purchasing power parity to ensure fair comparisons

1 Adjusting for Inflation



- As we have salaries across 2020-2023, it is important to bring them to one scale to make fair comparison
- We have scaled up the salary data from 2020 – 2022 to 2023 based on the ***inflation rates of each country***

2 Adjusting for Purchasing Power Parity



Country	Avg. DS Salary	Absolute Salary (USD)	PPP Adjusted Salary
India	₹ 1,300,000	$15.6K	**$56K**
China	¥293,000	$41.1K	**$70K**

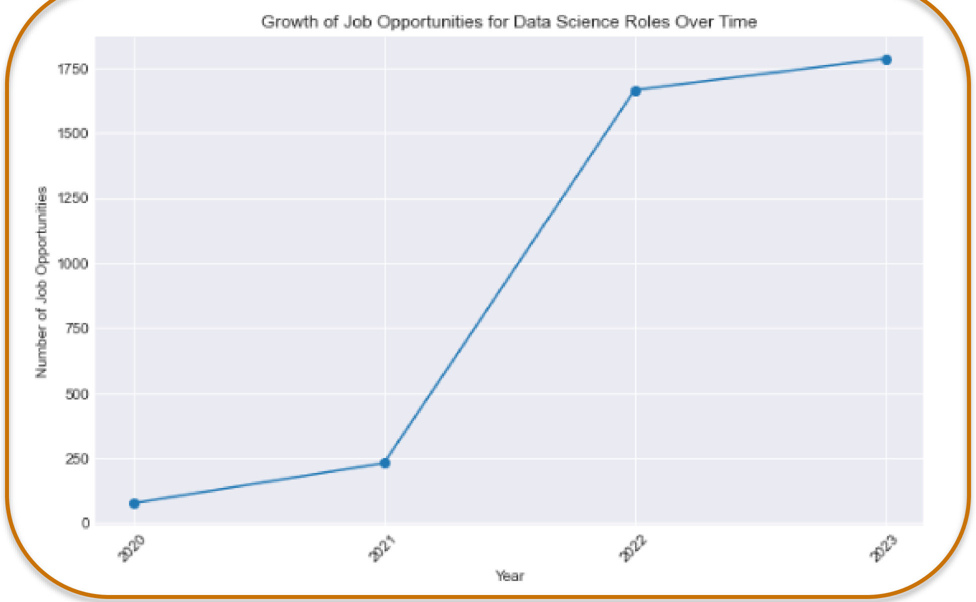
Normalized salary

AGENDA

- Background and Objectives
- Understanding the dataset and cleaning
- Exploratory data analysis
- Predictive Modeling
- Recommendations and conclusion

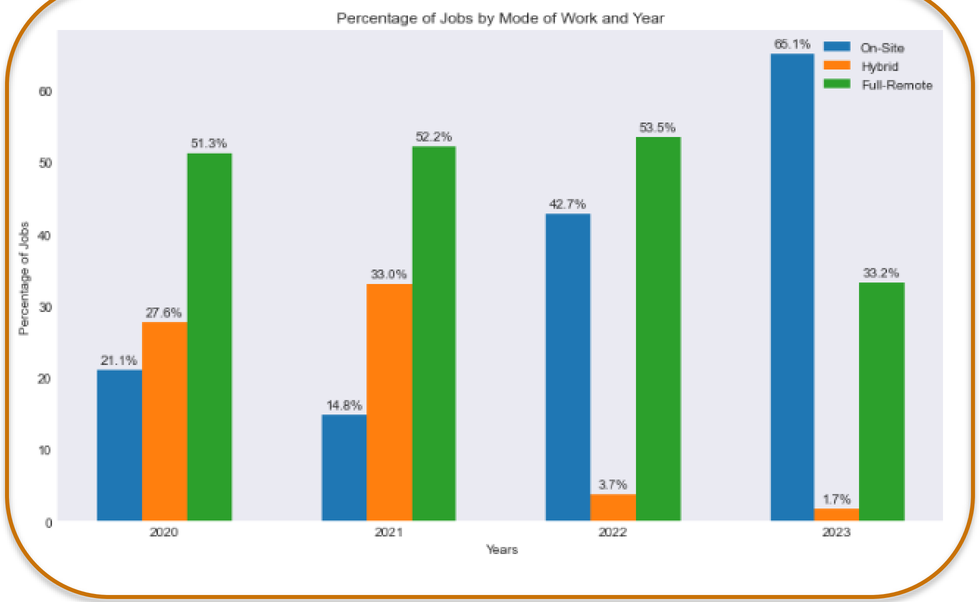
EDA – Opportunity Growth

Global Impact



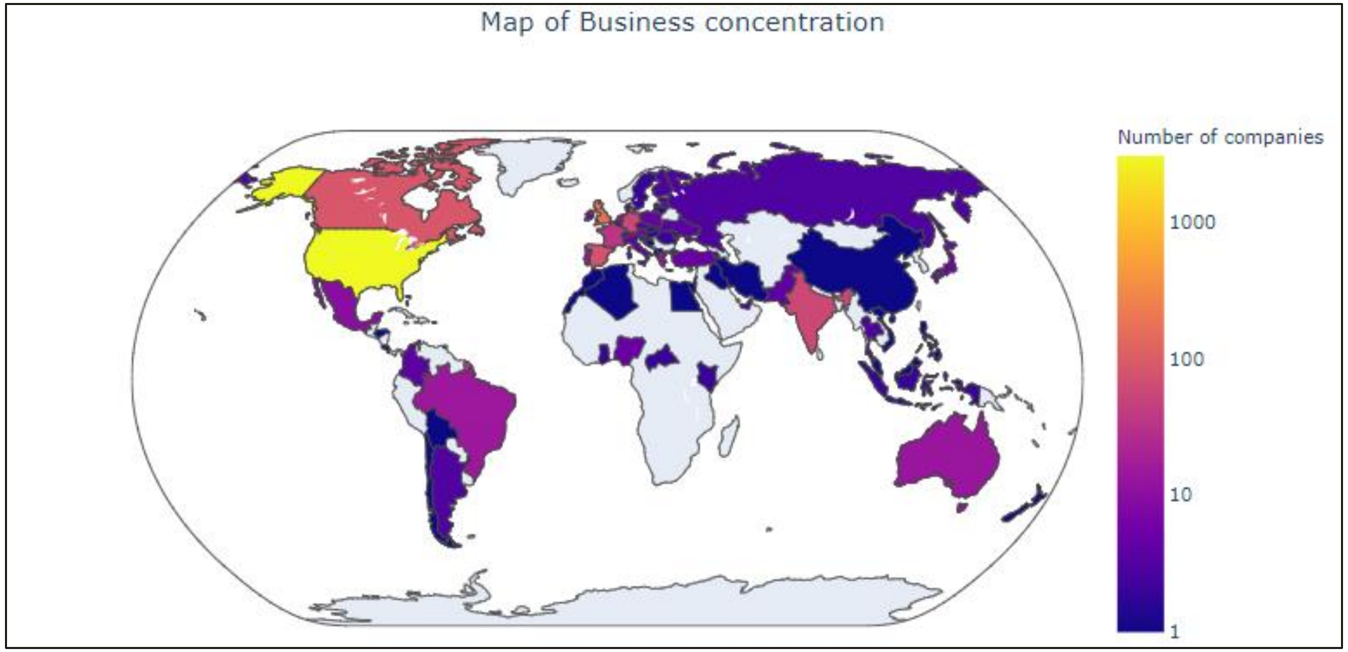
Over the years, data science opportunities have seen a remarkable surge across the globe.

Mode of Work



- Full-remote jobs decline from 2020 to 2023, shifting towards on-site opportunities indicating the **COVID impact**

EDA – Geographical Distribution

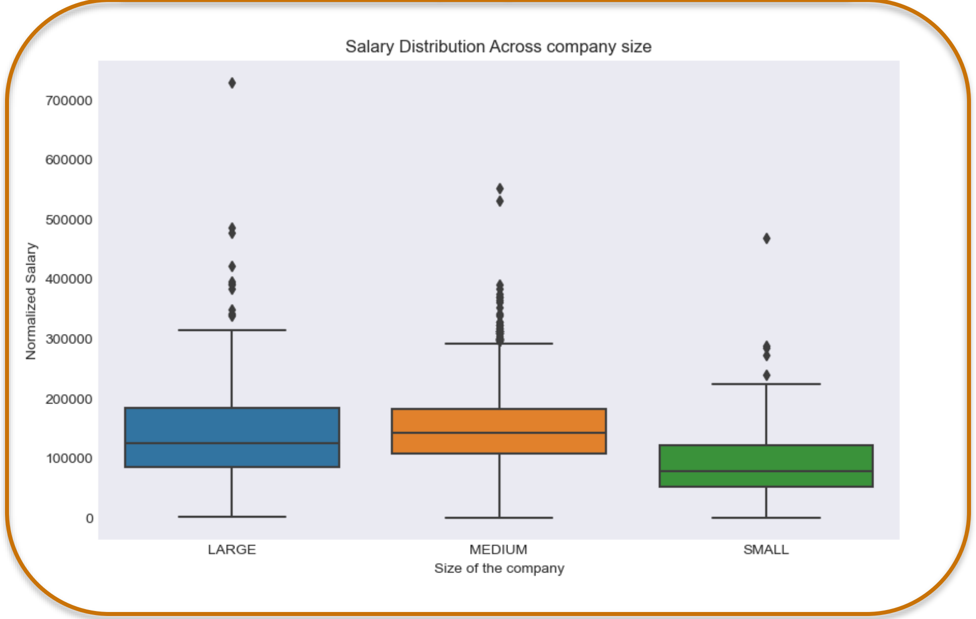


Key Takeaways

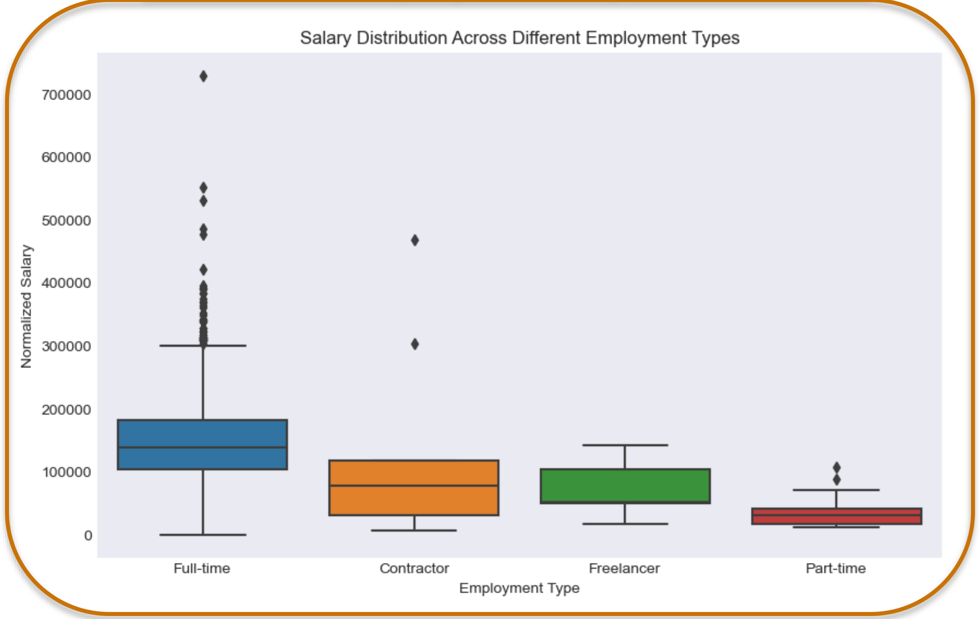
- ➢ Maximum data science roles lie around north American region.

EDA – Pay scale

Company Size



Employment Type

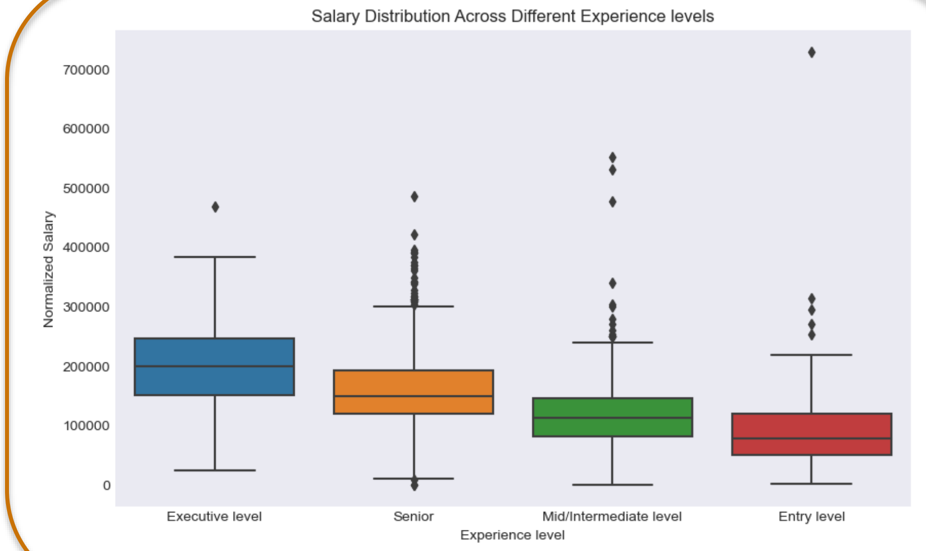


- ➢ Medium-sized companies offer higher average compensation than large or small-sized companies.

- ➢ Full-time employees enjoy a higher pay scale in comparison to other employee types.

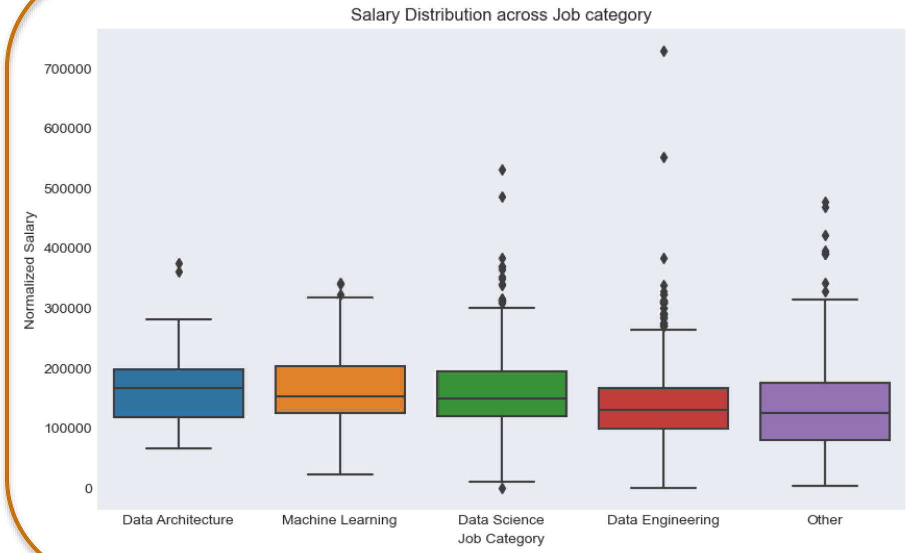
EDA – Pay scale

Experience Level



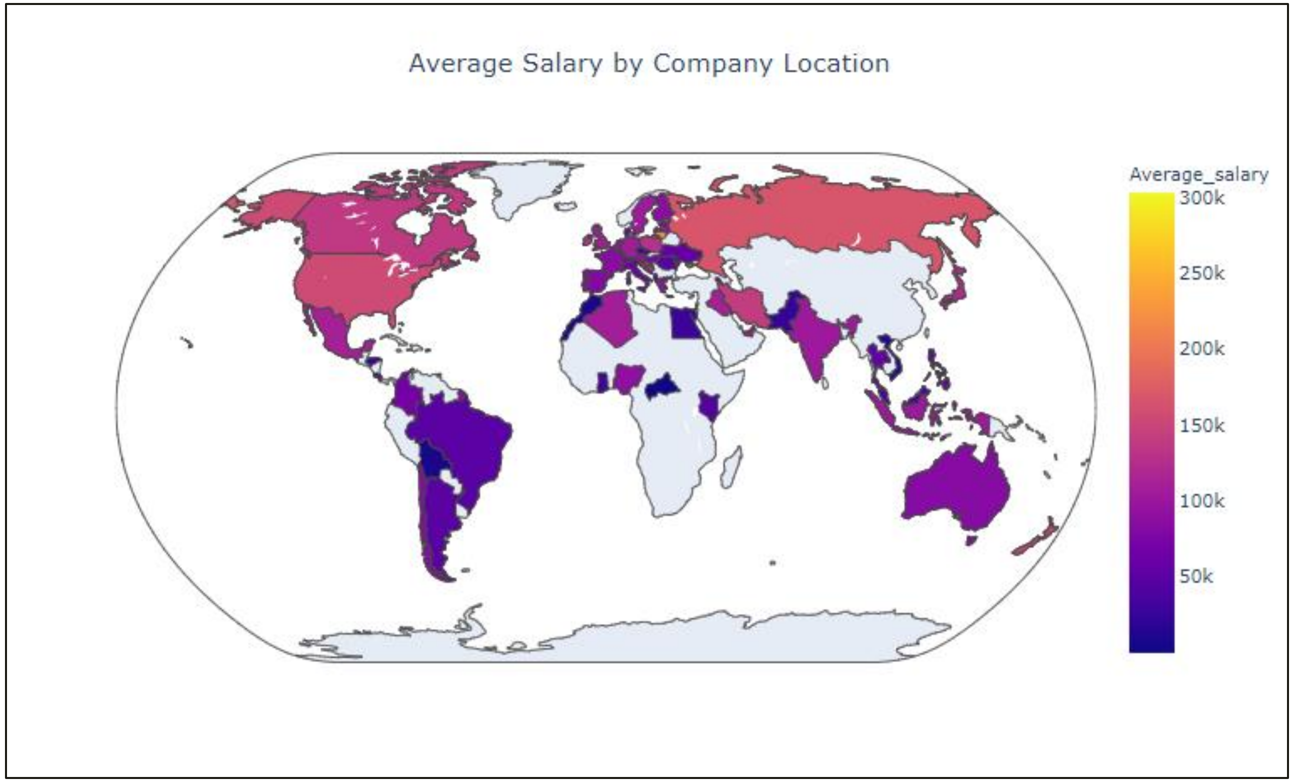
- As experience grows, so does the offered salary

Job Category



- Job roles do not significantly impact the Pay scale
- Among the most prevalent roles, data engineering positions exhibit comparatively lower salaries

EDA – Geographical Distribution



Key Takeaways

- Technologically developed countries such as the United States, Singapore, and Canada has the highest average salary.
- Less developed areas such as the majority of South America earned less.

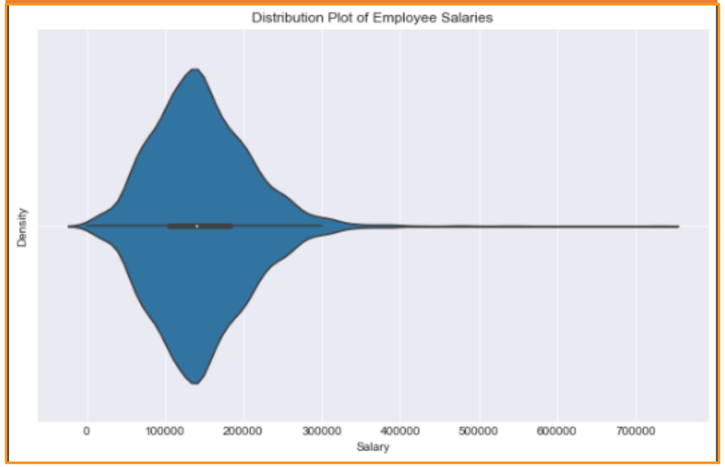
By analyzing these patterns, it is clear that the normalized salary metric has some relationship with the predictors (experience level, employment type, remote ratio, company size, company location and employee location)

AGENDA

- Background and Objectives
- Understanding the dataset and cleaning
- Exploratory data analysis
- Predictive Modeling
- Recommendations and conclusion

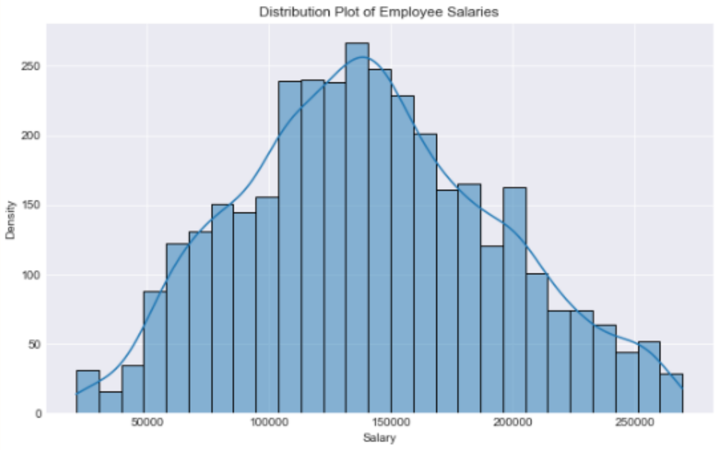
Data Preparation for Modelling

Identifying And Excluding Outliers



It is clear that there are a **few outliers** in the target variable. We will **remove records that fall outside of mean +/- 2 SD confidence interval** for the purpose of modelling (**~170 records**)

No Skew in salary



Salary is ***normally distributed***

Converting salary into category

- As our objective is to predict if the salary is going to fall above 75% of the people in this domain, we will be converting salary into a **binary categorical variable**
  - **1 --> Above 75th percentile**
  - **0 --> Below 75th percentile**

One Hot Encoding of Categorical Variables

In one hot encoding, we convert each categorical value of a variable into a **new categorical column and assign a binary value of 1 or 0** to those columns. *See example to the right:*

Mode of work
On-Site
Hybrid
Full-Remote

One Hot Encoding

On-Site	Hybrid	Full-Reote
1	0	0
0	1	0
0	0	1

Variable selection and handling class imbalance

<table>
<tr><th></th><th>Step 1</th><th>Step 2</th><th>Step 3</th><th>Step 4</th></tr>
<tr><td><b><i>Model →</i></b></td><td><b><i>Base Model</i></b></td><td><b><i>Lasso Regression</i></b></td><td><b><i>Model - After variable selection</i></b></td><td><b><i>Model - Class imbalance fixed</i></b></td></tr>
<tr><td><b><i>Description</i></b></td><td>A logistic regression model run with all the response variables</td><td>A Lasso regression model to identify the variable importance</td><td>Logistic regression model to see if there is improvement in prediction after variable selection</td><td>Updated the sampling technique based on <b><i>RandomUnderSampler</i></b> to fix the class imbalance</td></tr>
<tr><td><b><i>Results</i></b></td><td>➢ Accuracy: 76%
➢ Precision: 73%
➢ Recall: 76%</td><td>➢ Accuracy: 75%
➢ Precision: 70%
➢ Recall: 75%</td><td>➢ Accuracy: 76%
➢ Precision: 73%
➢ Recall: 76%</td><td>➢ Accuracy: 62%
➢ Precision: 73%
➢ Recall: 62%</td></tr>
<tr><td><b><i>Confusion Matrix</i></b></td>
<td><b><i>Training Data Split</i></b>
Below 75th → 1,891
Above 75th → 618
<table><tr><td></td><td></td><td colspan="2">Predicted</td></tr><tr><td></td><td></td><td>Below 75th</td><td>Above 75th</td></tr><tr><td rowspan="2">Actual</td><td>Below 75th</td><td>99%</td><td>1%</td></tr><tr><td>Above 75th</td><td>93%</td><td>7%</td></tr></table></td>
<td><b><i>Training Data Split</i></b>
Below 75th → 1,891
Above 75th → 618
<table><tr><td></td><td></td><td colspan="2">Predicted</td></tr><tr><td></td><td></td><td>Below 75th</td><td>Above 75th</td></tr><tr><td rowspan="2">Actual</td><td>Below 75th</td><td>99%</td><td>1%</td></tr><tr><td>Above 75th</td><td>97%</td><td>3%</td></tr></table></td>
<td><b><i>Training Data Split</i></b>
Below 75th → 1,891
Above 75th → 618
<table><tr><td></td><td></td><td colspan="2">Predicted</td></tr><tr><td></td><td></td><td>Below 75th</td><td>Above 75th</td></tr><tr><td rowspan="2">Actual</td><td>Below 75th</td><td>99%</td><td>1%</td></tr><tr><td>Above 75th</td><td>93%</td><td>7%</td></tr></table></td>
<td><b><i>Training Data Split</i></b>
Below 75th → 618
Above 75th → 618
<table><tr><td></td><td></td><td colspan="2">Predicted</td></tr><tr><td></td><td></td><td>Below 75th</td><td>Above 75th</td></tr><tr><td rowspan="2">Actual</td><td>Below 75th</td><td>60%</td><td>40%</td></tr><tr><td>Above 75th</td><td>30%</td><td>70%</td></tr></table></td></tr>
<tr><td><b><i>Takeaway</i></b></td><td>➢ <b><i>Very poor</i></b> Recall for Above 75th percentile due to low samples size for training – <b>High Class Imbalance</b></td><td>➢ <b>Removed</b> all the <b>variables</b> that were <b>penalized to 0 based on L1 penalty</b></td><td>➢ <b>No significant difference</b> before and after variable selection
➢ <b>Still High Class imbalance</b></td><td>➢ Though overall accuracy has dropped, the model does a <b>decent job in predicting both categories –</b> Class imbalance treated</td></tr>
</table>

Trying out different models to further improve prediction

<table>
<tr><th><em>Model →</em></th><th><em>KNN Classification</em></th><th><em>Random Forest</em></th><th><em>Boosting (GBM)</em></th><th><em>Neural Network (SGD)</em></th></tr>
<tr><td><em>Description</em></td><td>Identified the best K value using cross-validation and built a KNN classifier</td><td>Using 5 fold cross validation and Grid search, identified the best hyperparameters for RF</td><td>Using 5 fold cross validation and Grid search, identified the best hyperparameters for boosting</td><td>Using SGD built a neural network model</td></tr>
<tr><td><em>Results</em></td><td>➢ Best K: 27
➢ Accuracy: 76%
➢ Precision: 74%
➢ Recall: 76%</td><td>➢ Depth: 5 | Feature: 3 | Trees: 1000
➢ Accuracy: 62%
➢ Precision: 74%
➢ Recall: 62%</td><td>➢ λ=0.01| Depth: 3 | Trees: 500
➢ Accuracy: 62%
➢ Precision: 73%
➢ Recall: 62%</td><td>➢ Accuracy: 70%
➢ Precision: 73%
➢ Recall: 70%</td></tr>
<tr><td><em>Confusion Matrix</em></td>
<td><table><tr><th></th><th></th><th colspan="2">Predicted</th></tr><tr><th></th><th></th><th>Below 75th</th><th>Above 75th</th></tr><tr><td rowspan="2">Actual</td><td>Below 75th</td><td>98%</td><td>2%</td></tr><tr><td>Above 75th</td><td>87%</td><td>13%</td></tr></table></td>
<td><table><tr><th></th><th></th><th colspan="2">Predicted</th></tr><tr><th></th><th></th><th>Below 75th</th><th>Above 75th</th></tr><tr><td rowspan="2">Actual</td><td>Below 75th</td><td>60%</td><td>40%</td></tr><tr><td>Above 75th</td><td>30%</td><td>70%</td></tr></table></td>
<td><table><tr><th></th><th></th><th colspan="2">Predicted</th></tr><tr><th></th><th></th><th>Below 75th</th><th>Above 75th</th></tr><tr><td rowspan="2">Actual</td><td>Below 75th</td><td>60%</td><td>40%</td></tr><tr><td>Above 75th</td><td>30%</td><td>70%</td></tr></table></td>
<td><table><tr><th></th><th></th><th colspan="2">Predicted</th></tr><tr><th></th><th></th><th>Below 75th</th><th>Above 75th</th></tr><tr><td rowspan="2">Actual</td><td>Below 75th</td><td>75%</td><td>25%</td></tr><tr><td>Above 75th</td><td>45%</td><td>55%</td></tr></table></td></tr>
<tr><td><em>Takeaway</em></td><td>➢ <strong>Worse</strong> than Logistic Regression</td><td>➢ Same as Logistic Regression</td><td>➢ Same as Logistic Regression</td><td>➢ <strong>Slightly better than logistic regression / RF/ Boosting</strong> due to better accuracy and similar recall%</td></tr>
</table>

AGENDA

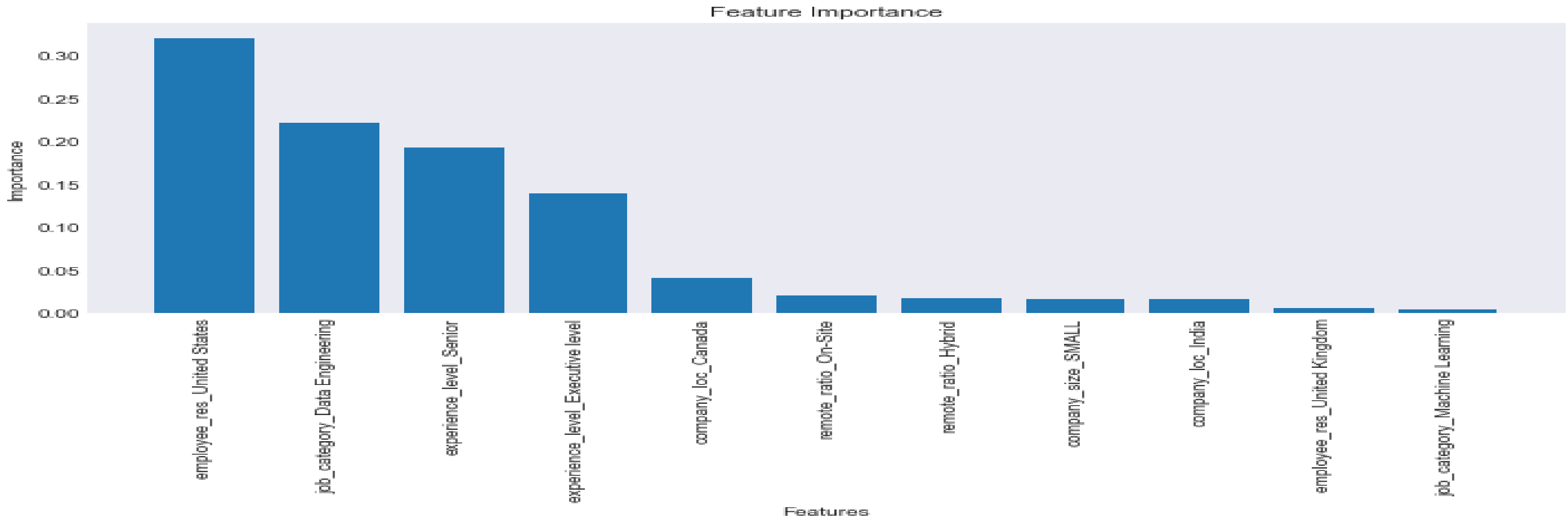
- Background and Objectives
- Understanding the dataset and cleaning
- Exploratory data analysis
- Predictive Modeling
- Recommendations and conclusion

Recommendations and conclusion

<table>
<tr><th><em>Model →</em></th><th>Logistic Regression / RF / Boosting</th><th><em>Neural Network (SGD)</em></th></tr>
<tr><td><strong><em>Results</em></strong></td><td>➢ Accuracy: 62%
➢ Precision: 73%
➢ Recall: 62%</td><td>➢ Accuracy: 70%
➢ Precision: 73%
➢ Recall: 70%</td></tr>
<tr><td><strong><em>Confusion Matrix</em></strong></td>
<td><table>
<tr><th></th><th></th><th colspan="2">Predicted</th></tr>
<tr><th></th><th></th><th>Below 75th</th><th>Above 75th</th></tr>
<tr><th rowspan="2">Actual</th><th>Below 75th</th><td>482 (60%)</td><td>322 (40%)</td></tr>
<tr><th>Above 75th</th><td>83 (30%)</td><td>189 (70%)</td></tr>
</table></td>
<td><table>
<tr><th></th><th></th><th colspan="2">Predicted</th></tr>
<tr><th></th><th></th><th>Below 75th</th><th>Above 75th</th></tr>
<tr><th rowspan="2">Actual</th><th>Below 75th</th><td>603 (75%)</td><td>201 (25%)</td></tr>
<tr><th>Above 75th</th><td>122 (45%)</td><td>150 (55%)</td></tr>
</table></td></tr>
</table>

Recommendations and conclusions

- ➢ While the neural network model performed better than other models, the prediction accuracy, precision and recall are not so high
- ➢ Maybe advanced feature engineering techniques and a bigger sample size can help predict the salary better


Feature Importance
Importance
employee_res_United States
job_category_Data Engineering
experience_level_Senior
experience_level_Executive level
company_loc_Canada
remote_ratio_On-Site
remote_ratio_Hybrid
company_size_SMALL
company_loc_India
employee_res_United Kingdom
job_category_Machine Learning
Features