SC1015 Mini Project

FCSG Team 1

Goh Jun Keat	U2320114E
Celeste Ho Shir Chee	U2322765G
He QiXin	U2321190F

Table of contents

01

Motivation

Are Data Science Professionals getting paid above the median salary (USD)?

02

Data Cleaning

Preparation of Dataset

03

EDA

Breakdown & Analysis of Variables

04

Models

- Binary Classification
  - Random Forest
- Logistic Regression

05

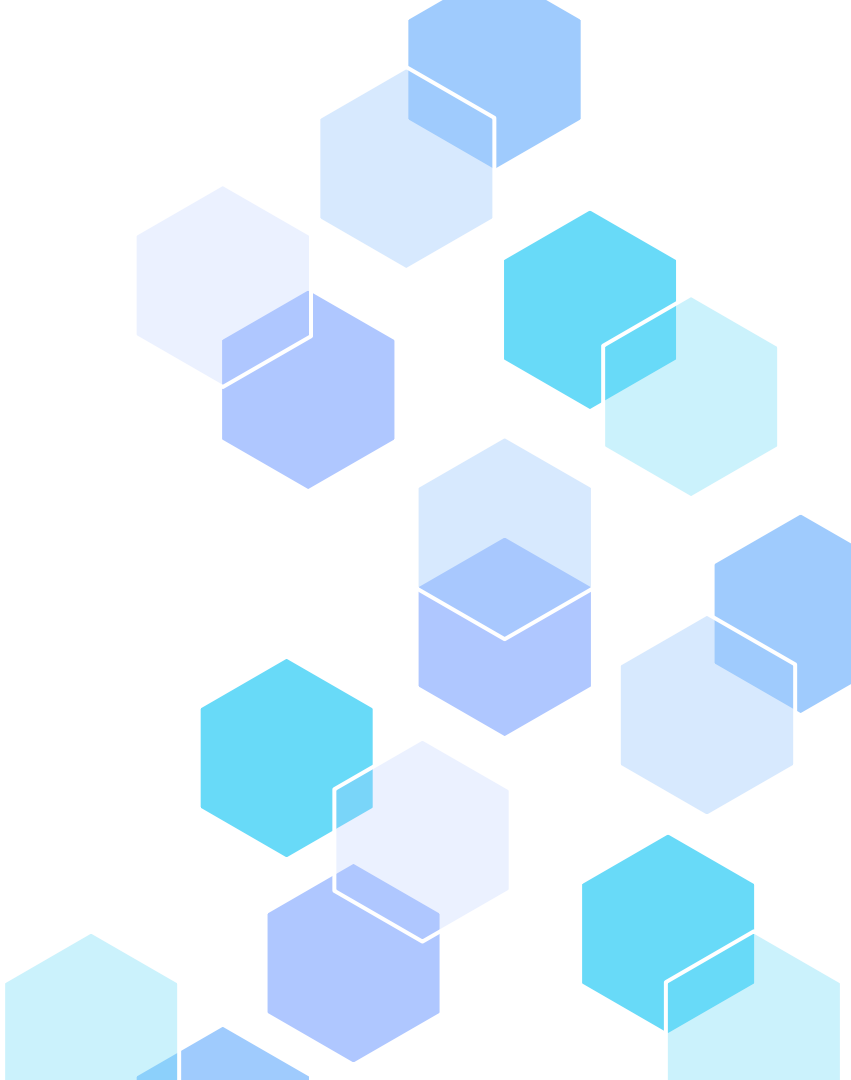
Findings

Conclusion &
Data-Driven Insights

01

Motivations

Data Science Salaries



Motivations

- The field of Data Science is increasingly popular in this digitalised world.
- More people are interested in becoming Data Science professionals.
- Data Science remains a relatively new field with many uncertainties:
    - Uncertainty about job opportunities.
    - Few potential work locations
    - Some Data Science jobs pay more than others

- Aim:
    - To investigate why some data science professionals are **getting paid more than others**
    - To gather **more insights** into Data Science related professions

Dataset

- Dataset used: **“Data Science Salaries”** by Zain Faisal

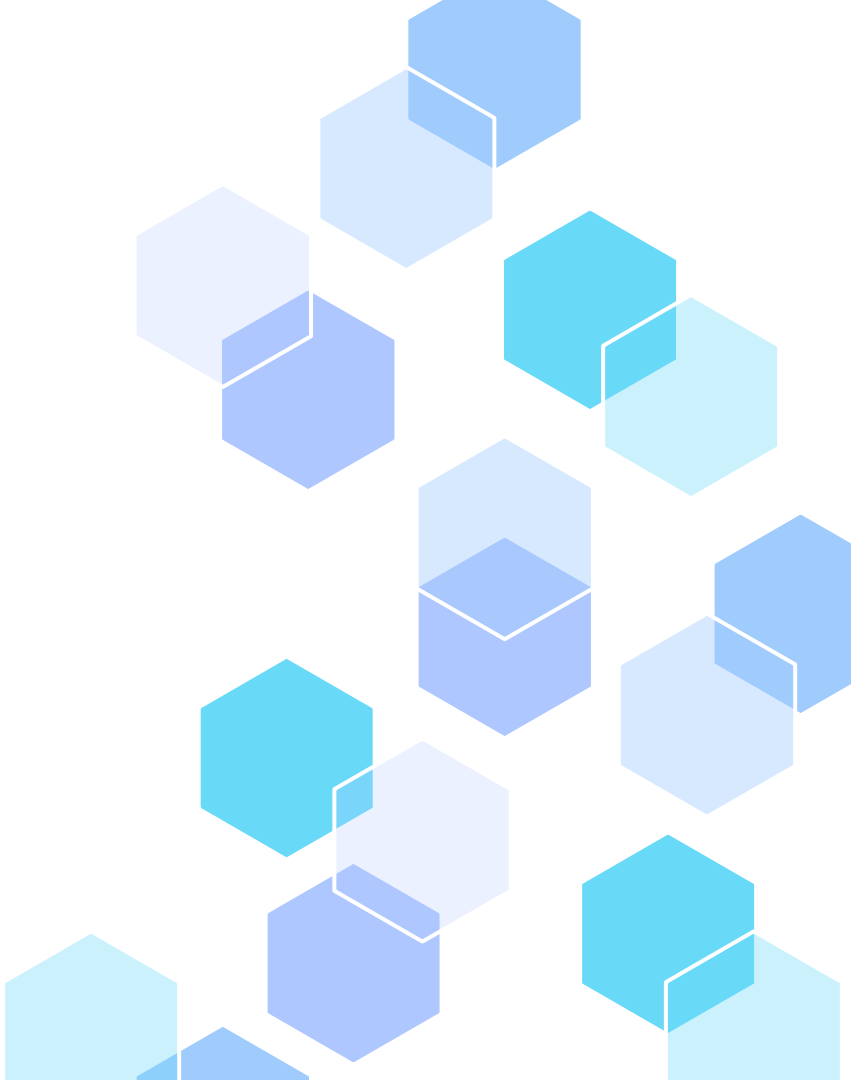
- Used data of Data Science Professionals from 3 years:
  - Total of 606 respondents across 2020, 2021 and 2022.

- Dataset includes 12 columns based on the respondent’s profile:
  - 9 columns contribute to the respondent’s salary and salary (USD).

02

Data Cleaning

Preparation of Dataset



Data Cleaning

- Checked for missing values and duplicates.

- Removal of Unnecessary Columns:
  - “id”, “salary”, “salary_currency”, “employee_residence”
  - For standardisation of salaries in USD.

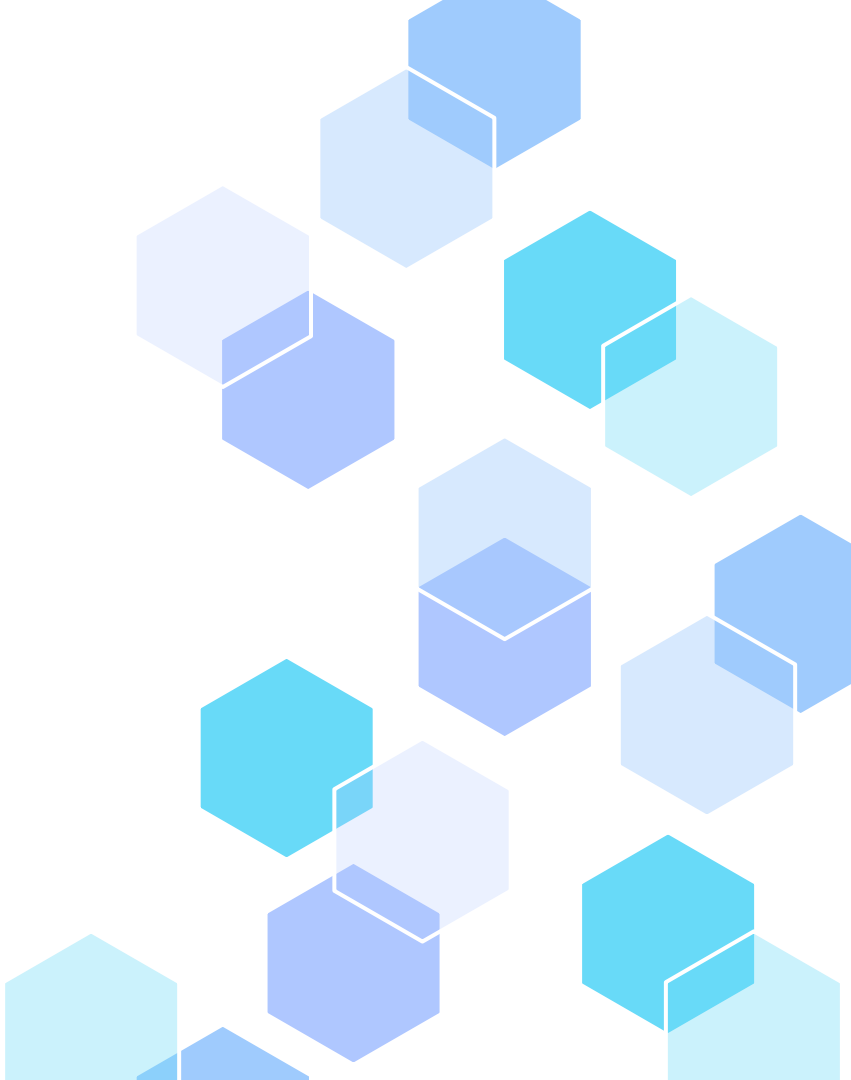
- Renaming of Values in Columns:
  - For better understanding of the data

- Addition of New Columns:
  - “job_category” to categorise the many different Data Science jobs.
  - **“above_median” to know if respondent has above median salary (USD).**

03

Exploratory Data Analysis

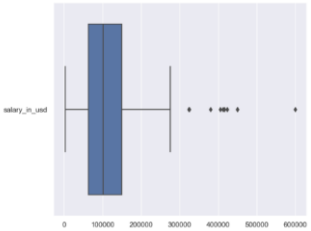
Breakdown & Analysis after Data Cleaning

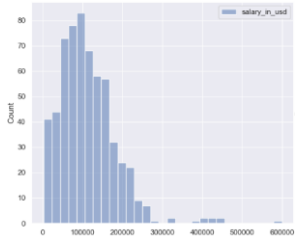


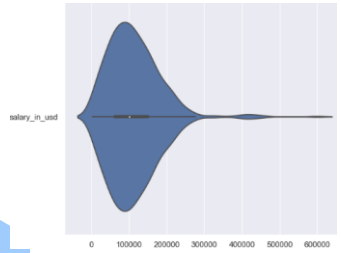
Breakdown of Variables

Numeric

- 2 numeric variables
- Only 1 truly considered numeric
- Used **box-plot, violin-plot, histogram.**

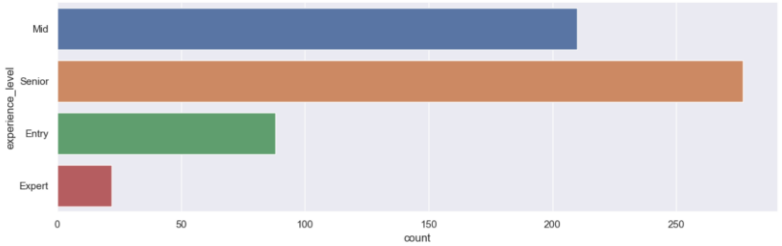


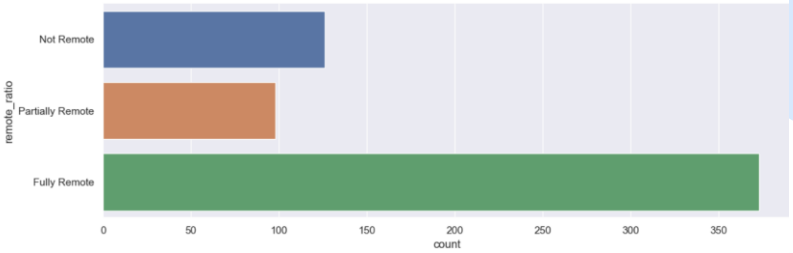




Categorical

- 7 categorical variables
- Used **catplot** to plot distribution of each variable

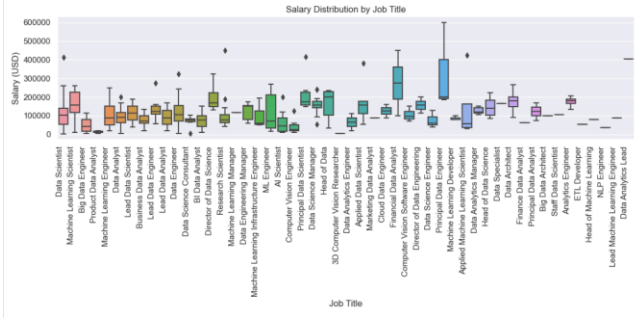




Analysis of Variables I

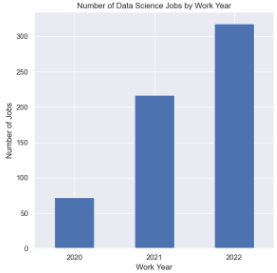
Boxplot

- Distribution of salary (USD) by job title



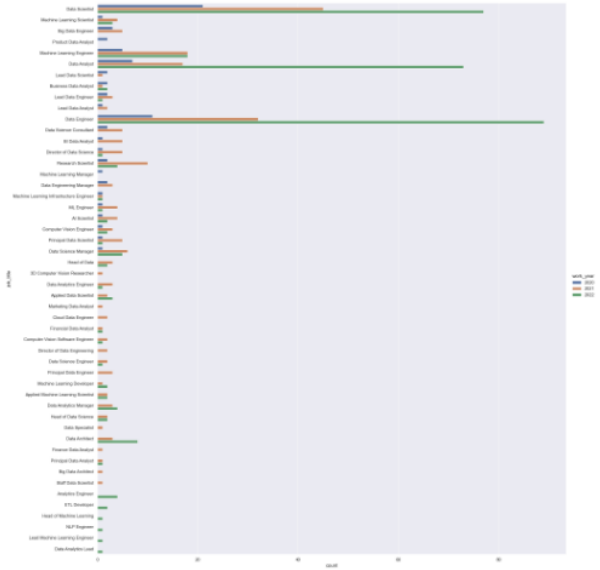
Barplot

- Top 10 data science company locations
- Trend of job opportunities by year



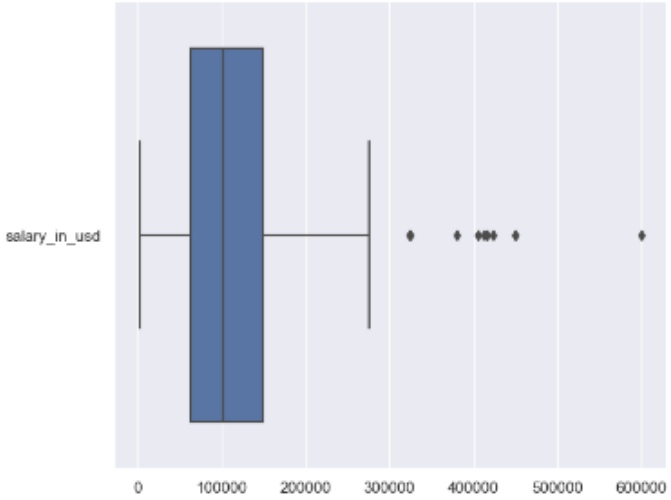
Catplot

- Most popular job title & job category by year



Outliers

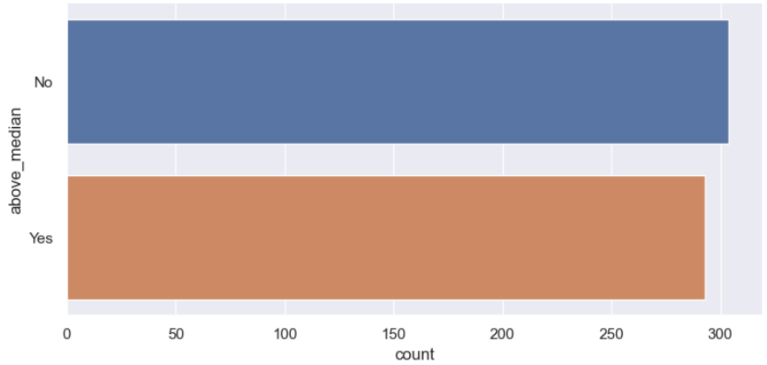
- Variable “salary_in_usd” has outliers that can be detected by boxplot.
- Outliers were dropped to increase reliability
  - 10 rows of data were dropped



Analysis of Variables II

Catplot

- Distribution of responding variable “above_median”



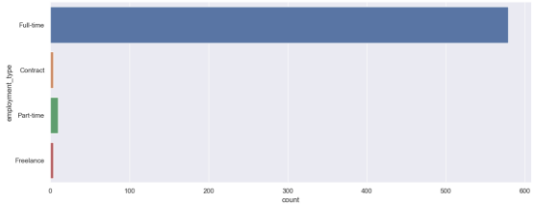
Line Graph

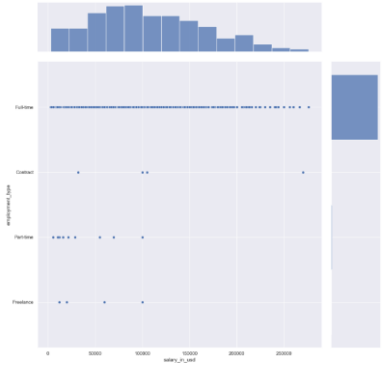
- Trend of average salary (USD) by year

Median value used is the median value from **raw data (before removing outliers)**

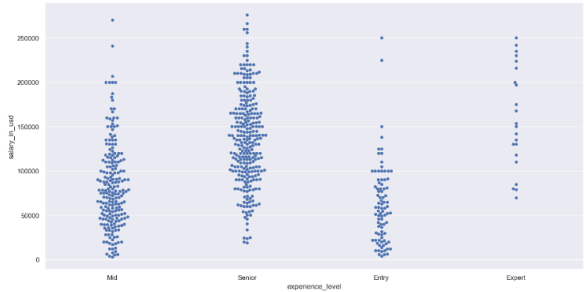
Predictors

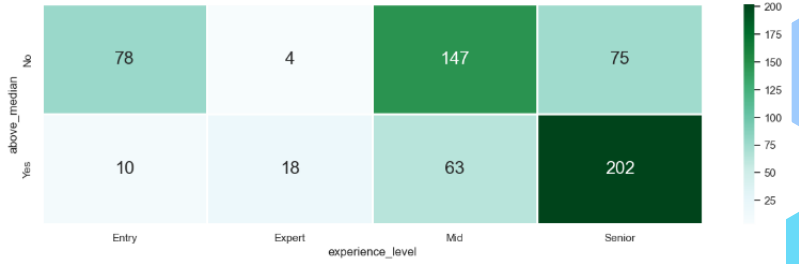
- 6 **potential** predictors for “above_median”
- Used catplot and jointplot to identify **valid predictors.**





- **3 valid predictors found. (experience_level, remote_ratio, company_size)**
- Used swarmplot to find distribution of **valid predictors** with salary (USD).
- Categorise **valid predictors** against “above_median” using heatmap and GroupBy.





04

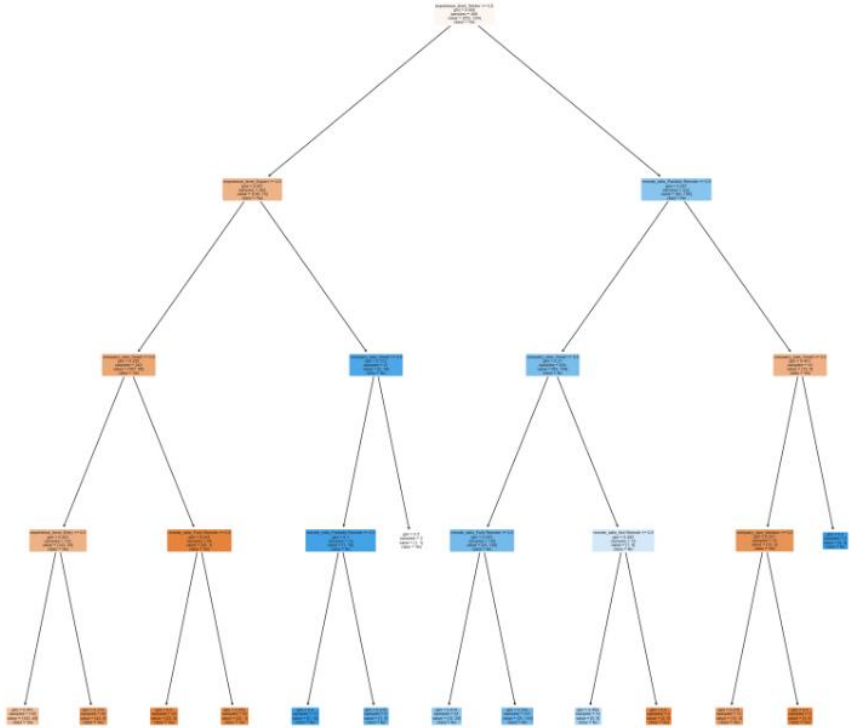
Models

Prediction Accuracy

Binary Classification

- Encoded the **valid predictors** by One-Hot Encoding.
- Dataset split into Train and Test by ratio 8:2
- Max depth set to 4
- Upsampling done to increase accuracy

- Analysis:
  - Accuracy: **72.95%**
  - False Positive Rate (FPR): **21.15%**
  - False Negative Rate (FNR): **31.43%**

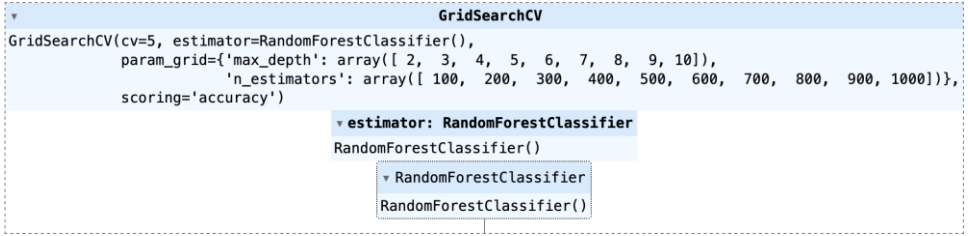


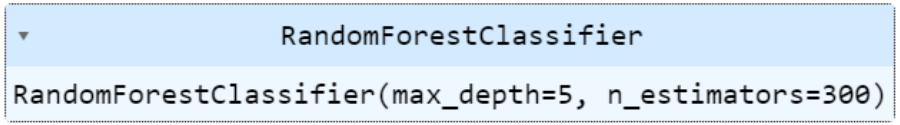
Random Forest

- The dataset was split into Train and Test by ratio of 8:2.

- 300 decision trees with depth 5 (chosen via hyper-paramater tuning using Cross-Validation (CV), using **accuracy** as the scoring parameter)

- Analysis
  - Accuracy: **81.97%**
  - False Positive Rate (FPR): **17.31%**
  - False Negative Rate (FNR): **18.57%**





Logistic Regression

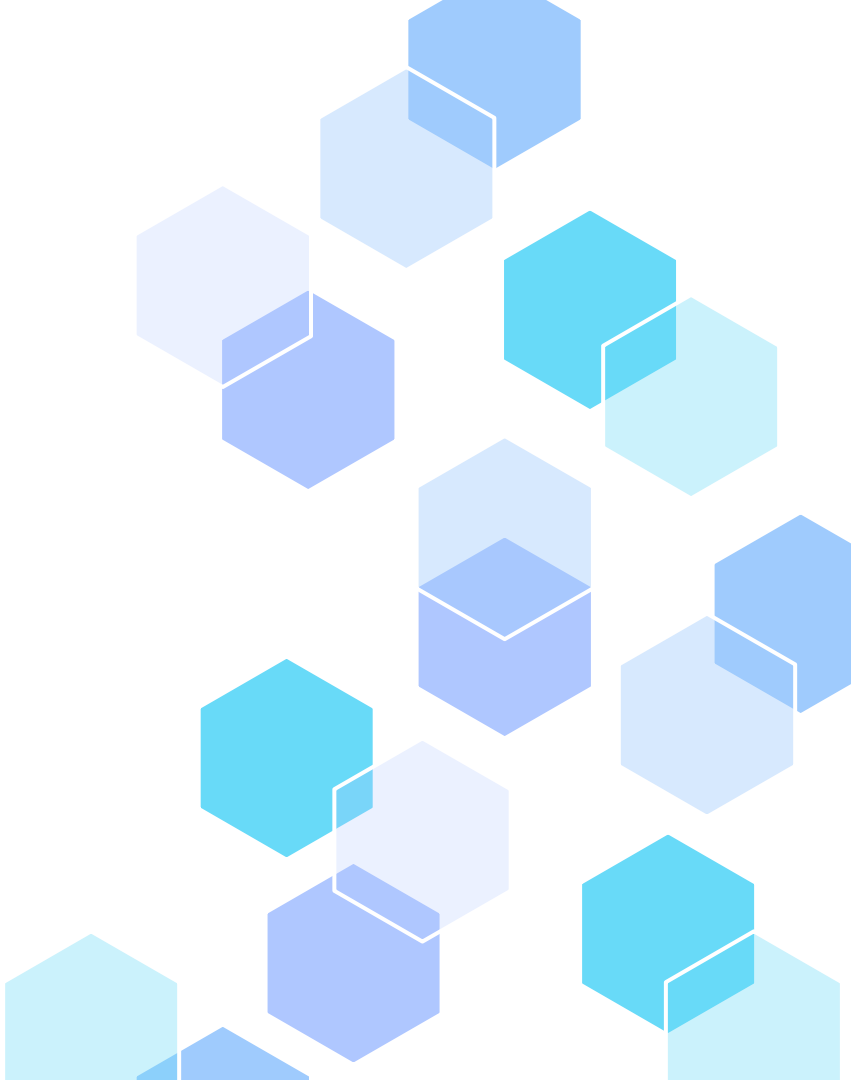
- The dataset was split into Train and Test by ratio of 8:2.
- After training, the code uses the trained model to make predictions on the testing data using “predict()”.
- Analysis of Test data:
  - Accuracy: 75.41%
  - False Positive Rate (FPR): 44.83%
  - False Negative Rate (FNR): 37.50%

```
Accuracy: 0.7540983606557377
Classification Report:
              precision    recall  f1-score   support

          No       0.73      0.78      0.75        58
         Yes       0.78      0.73      0.76        64

    accuracy                           0.75       122
   macro avg       0.75      0.76      0.75       122
weighted avg       0.76      0.75      0.75       122
```

05

Findings

Conclusion & Data-Driven Insights



Conclusion

Binary Classification

Valid predictors can predict whether salary (USD) of respondent is above median

Random Forest

Valid predictors can predict whether salary (USD) is above median and has the **lowest false positive and false negative rates.**

Logistic Regression

Valid predictors can predict whether salary (USD) is above median salary (USD). This model has the highest false positive and false negative rates out of the 3 models.

Random Forest is the more suitable model to predict whether salary (USD) of Data Science professional is above the median salary (USD) as it has higher accuracy.

Data-Driven Insights

- Different Data Science jobs have different distribution of salary (USD).
  - **Financial Data Analysts** have the **highest median salary (USD).**
  - **Principle Data Engineers** have the **highest minimum salary (USD).**
- **Most popular job** in:
  - 2020 & 2021: **Data Scientist**
  - 2022: **Data Engineer**
- **Most popular job category** in:
  - 2020 & 2021: **Data Scientist**
  - 2022: **Data Analyst**
- Most Data Science-related companies are found in the **USA**.
- The **average salary (USD)** of Data Science professionals is **increasing by year.**
- There are **increasing number of job opportunities** in the Data Science field by year.

Thank you

References

https://stackoverflow.com/questions/57165247/rename-column-values-using-pandas-dataframe

https://www.dataquest.io/blog/tutorial-add-column-pandas-dataframe-based-on-if-else-condition/

https://www.geeksforgeeks.org/how-to-customize-line-graph-in-jupyter-notebook/

https://www.w3schools.com/python/python_ml_logistic_regression.asp

https://realpython.com/logistic-regression-python/

https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

https://medium.com/@24littledino/accuracy-in-python-980074154e52