CS989: Big Data Fundamentals Report

"Salary Prediction Classification"

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Salary has been used as compensation given to employees for achieving the companies' goals and targets (Didit and Nikmah, 2020). High salaries have always been a pivotal factor when choosing which industry and job role for many individuals. However, as Millennials and Generation Z take over as the predominant working class, the demand for higher salaries increases (Alonso-Almeida and Llach, 2019). This phenomenon may also derive from the soaring global inflation rate (Ozili and Arun, 2023), making individuals lose purchasing power and urge for increased compensation to satisfy their needs (Asaari, Desa and Subramaniam, 2019). Hence, this study aims at uncovering patterns that highlight the pivotal attributes and competencies correlated with higher income.

This report will examine data extracted by Barry Becker from the 1994 Census database with over 40 countries. Moreover, with the diverse range of features the goal is to have a holistic approach to our research.

Chapter 2

Dataset - *"Salary Prediction Classification"*

Aims

For this assignment, we will analyse a dataset with different salaries from different countries. Our goal is to compare the salaries in the UK to the salaries of the US and implement a model that will accurately predict salaries.

Source

The dataset is publicly available on Kaggle and its original source is the UC Irvine Machine Learning Repository. Find below the two links to the dataset:

https://www.kaggle.com/datasets/ayessa/salary-prediction-classification/data
https://archive.ics.uci.edu/dataset/20/census+income

The dataset exists in only one .csv sheet. Since the data is from the year 1994, the insights may have lost relevancy value. However, we believe that insights drawn from this dataset may help understanding existing salary trends. Nonetheless, we choose this dataset for the reasons below:

- With a total of 15 features, the dataset is balanced, ensuring that the analysis generates relevant results within the timeframe.
- There are 32561 rows, which will not require sampling or special hardware to produce analytical results.
- Contains data from different countries, allowing a cross-cultural comparison to be made.
- The data exists in one place, therefore, partaking in single-source-of-truth (SSOT) activities is not required.

Description

The dataset is described by both Kaggle and the UC Irvine Machine Learning Repository website. Additionally, a description of the features will be conducted, imprinting our understanding resulting from the analysis conducted. The following table describes the features in the dataset and the possible values they could have:

Table 2.1: Features of The Dataset.

Feature Name	Type	Possible Values
age	Integer	Discrete Integer Values.
workclass	Categorical	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt	Integer	Integer Values.
education	Categorical	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num	Integer	Integer Values.
marital-status	Categorical	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation	Categorical	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship	Categorical	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race	Categorical	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex	Binary	Female, Male.
capital-gain	Integer	Integer Values.

capital-loss	Integer	Integer Values.
hours-per-week	Integer	Integer Values.
native-country	Categorical	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
salary	Binary	>50K, <=50K.

Analysis

To grasp the dataset, exploratory data analysis will be executed. Filtering unnecessary features helps to gather insights for developing a suitable. This dataset exploration further supports a wider understanding of future challenges once transforming and cleaning the data to create the machine learning model.

Chapter 3

Exploratory Data Analysis and Data Preprocessing

Exploratory Data Analysis (EDA)

The dataset contains 15 columns and 32,561 rows, indicating more than 480,000 values. Most values are of type object meaning that later we will have to encode the values so that they can be fed into a machine learning algorithm. Presently, our strategy was to do:

- General overview of the columns and their values.
- Check for any common problems.
- Compare the UK and other countries.
- If any problems arise, they will be handled later.

Starting with the first step of the analysis, overviewing the dataset by using arithmetic operations such as the mean, median, mode, etc. We identified that the average age is 38 years and the average working hours was 40 hours. This type of analysis is useful for the integer features, but rarely useful when it comes to categorical features.

To better analyse the categorical features, we needed to understand what the unique values for these features were. Once completed, the first problem emerged; a whitespace at the beginning of each categorical value, seen in Figure 3.1.

```
The unique values of column workclass are: [' State-gov' ' Self-emp-not-inc' ' Private' ' Federal-gov' ' Local-gov'
 ' ?' ' Self-emp-inc' ' Without-pay' ' Never-worked']
The unique values of column education are: [' Bachelors' ' HS-grad' ' 11th' ' Masters' ' 9th' ' Some-college'
 ' Assoc-acdm' ' Assoc-voc' ' 7th-8th' ' Doctorate' ' Prof-school'
 ' 5th-6th' ' 10th' ' 1st-4th' ' Preschool' ' 12th']
The unique values of column marital-status are: [' Never-married' ' Married-civ-spouse' ' Divorced'
 ' Married-spouse-absent' ' Separated' ' Married-AF-spouse' ' Widowed']
The unique values of column occupation are: [' Adm-clerical' ' Exec-managerial' ' Handlers-cleaners' ' Prof-specialty'
 ' Other-service' ' Sales' ' Craft-repair' ' Transport-moving'
 ' Farming-fishing' ' Machine-op-inspct' ' Tech-support' ' ?'
 ' Protective-serv' ' Armed-Forces' ' Priv-house-serv']
The unique values of column relationship are: [' Not-in-family' ' Husband' ' Wife' ' Own-child' ' Unmarried'
 ' Other-relative']
The unique values of column race are: [' White' ' Black' ' Asian-Pac-Islander' ' Amer-Indian-Eskimo' ' Other']
The unique values of column sex are: [' Male' ' Female']
The unique values of column native-country are: [' United-States' ' Cuba' ' Jamaica' ' India' ' ?' ' Mexico' ' South'
 ' Puerto-Rico' ' Honduras' ' England' ' Canada' ' Germany' ' Iran'
 ' Philippines' ' Italy' ' Poland' ' Columbia' ' Cambodia' ' Thailand'
 ' Ecuador' ' Laos' ' Taiwan' ' Haiti' ' Portugal' ' Dominican-Republic'
 ' El-Salvador' ' France' ' Guatemala' ' China' ' Japan' ' Yugoslavia'
 ' Peru' ' Outlying-US(Guam-USVI-etc)' ' Scotland' ' Trinadad&Tobago'
 ' Greece' ' Nicaragua' ' Vietnam' ' Hong' ' Ireland' ' Hungary'
 ' Holand-Netherlands']
The unique values of column salary are: [' <=50K' ' >50K']
```

Figure 3.1: The Whitespace Problem in The Categorical Values.

This problem will be handled in the preprocessing phase. However, when reviewing the categorical values, we uncovered a second problem; an existing '?' value in three columns: *workclass*, *occupation*, and *native-country*. A count of the '?' value in the three features was done to assess the severity of the problem, depicted in Figure 3.2.

```
[9]  df['workclass'].loc[df['workclass'] == ' ?'].count()

     1836

[10] df['occupation'].loc[df['occupation'] == ' ?'].count()

     1843

[11] df['native-country'].loc[df['native-country'] == ' ?'].count()

     583
```

Figure 3.2: Count of The '?' Value in *workclass*, *occupation*, and *native-country*.

The number of rows containing the '?' values for the three columns were extensive. Removing the rows containing these values would negatively impact the

results of the analysis. Within the rows containing the '?' value, it was found that for every *workclass* feature that had the value, there would be a corresponding value in the *occupation* feature. The same could not be said for the *occupation* feature nor the *native-country* feature. Later in the data preprocessing phase these columns would be handled uniquely.

Afterwards, we reviewed other common problems in the dataset. A common problem is defined to be an issue that exists in most datasets, including rows with null values, duplicated rows, inconsistent values, etc. Thus, we began analysing for any null values. There were no null values and we double-checked as it is a common problem. Next, we discovered 24 rows that had duplicate values. This meant that the problem was negligible and we could drop the rows without heavily impacting our analysis.

Subsequently, we inspected the dataset for any other inconsistencies, usually mismatched values in the dataset. Whilst no inconsistencies were found, two anomalies were discovered in the *native-country* column. The first anomaly is the *native-country* had 'Yugoslavia' values, which represented an eastern-European country that ceased to exist in 1992 and this dataset is from 1994. The second anomaly derives from the highest count in the *native-country* column is the 'United-States' value at 29,170 values, therefore accounting more than 89% of the dataset, as showcased in Figure 3.3.

```
[ ]  df['native-country'].value_counts()
```

United-States	29170
Mexico	643
?	583
Philippines	198
Germany	137
Canada	121
Puerto-Rico	114
El-Salvador	106
India	100
Cuba	95
England	90
Jamaica	81
South	80
China	75
Italy	73
Dominican-Republic	70

Figure 3.3: Count of The ‘United-States’ Value in *native-country* Compared to Others.

This suggests that we are handling an imbalanced dataset, as it includes worker and salary information from varied countries, including the UK as ‘England’ and ‘Scotland’, but most of the information is regarding the US.

We proceeded by comparing the UK and the US since both became singularly relevant to us. As aforementioned, the ‘UK’ value did not exist and the UK in the dataset was represented by ‘England’ and ‘Scotland’. All rows with those two values were combined into one UK dataset whilst all rows containing the ‘United-States’ value were combined into one US dataset. We extracted insights from the two datasets; the mean age value for the UK dataset was 41 years while for the US dataset was 38 years. The most frequent educational level in both datasets was High School. In both datasets, most people have worked in the private sector in a managerial or an executive role.

Although these insights can be gathered and understood by practitioners, it is considered a best-practice to utilise graphical data visualisation when presenting this information publicly. Before proceeding with the visualisation of the most important

differentiating points between the two countries, we decided to focus on two factors: correlation and the existence of outliers in the dataset, portrayed in Figure 3.4.

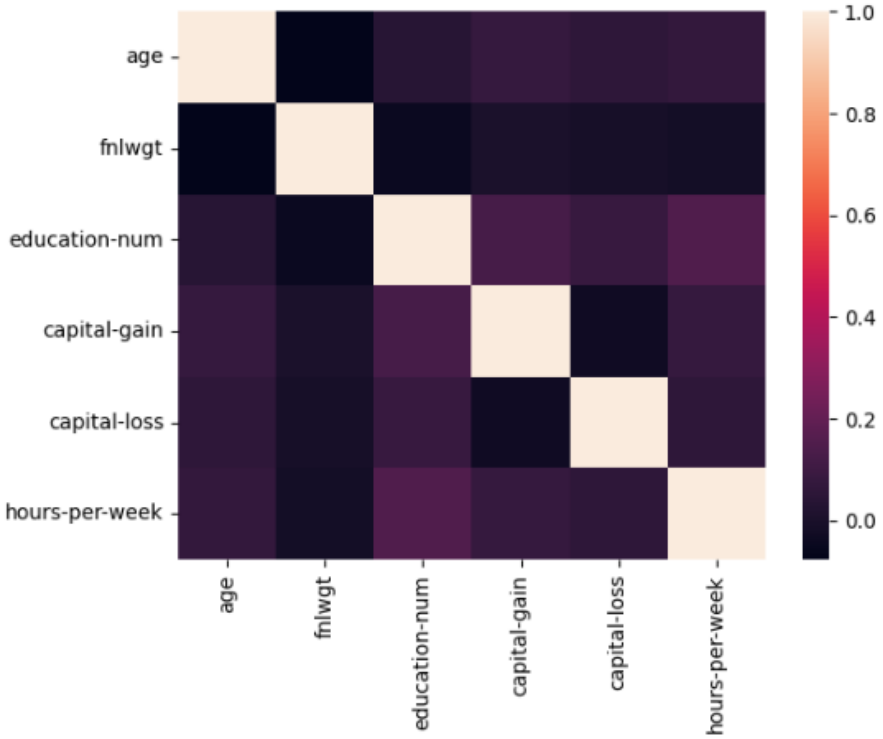

Figure 3.4: The Correlation Matrix.

The matrix suggests no strong correlation between any of the features. This signifies that any changes in the values of any features would not result in any strong change in any other features. Any movement in that direction was stopped. Next, we explored for any outliers in the dataset, specifically in the non-categorical features, as seen in Figure 3.5.

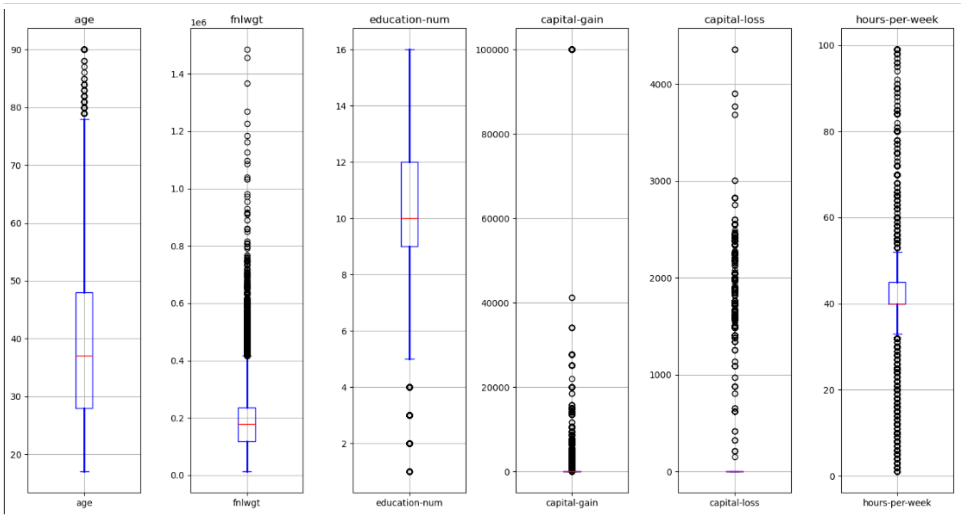

Figure 3.5: The Box Plots of Integer Features.

Due to the abundance of outliers in the dataset, it would not be wise to deal with them by removing or manipulating them as it would impact the analytical results.

We re-focused our efforts on comparing the UK and the US using graphs. We inspected the age ranges, and the salary ranges in both UK and US individuals, as showcased by Figures 3.6 and 3.7, respectively.

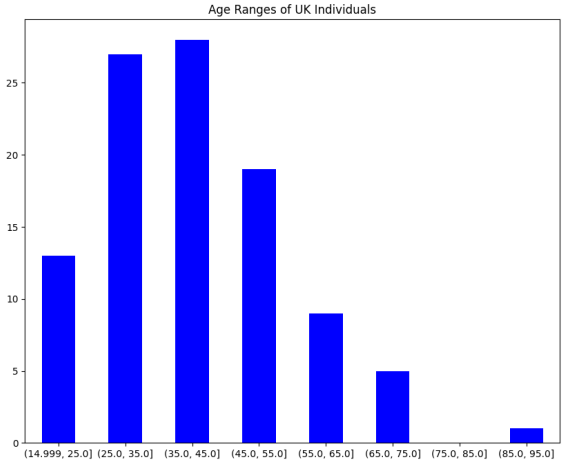

Figure 3.6: Age Ranges of UK Individuals.

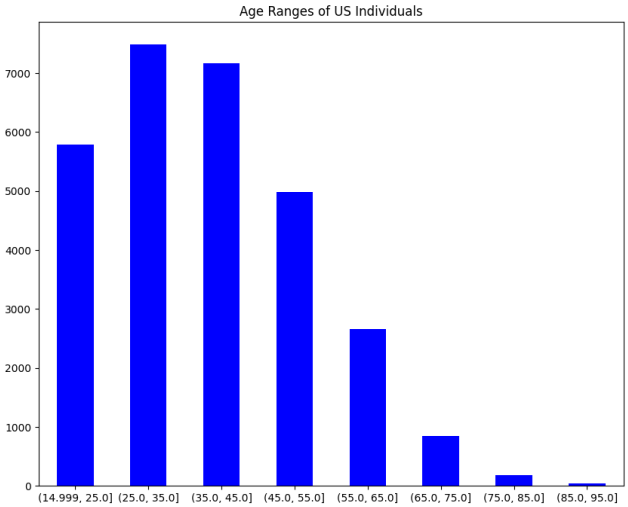

Figure 3.7: Age Ranges of US Individuals.

Considering their similarity, there is a positive correlation between the two datasets. Afterwards, a comparison of the salary ranges between UK and US individuals, displayed in Figures 3.8 and 3.9, respectively.

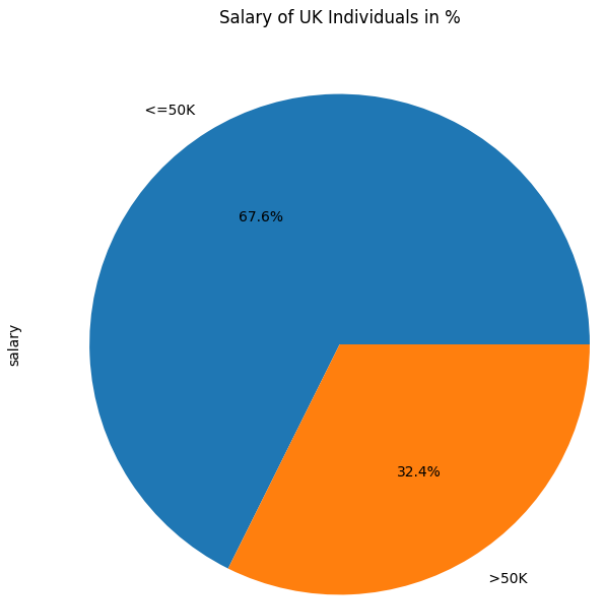

Figure 3.8: Salary of UK Individuals in %.

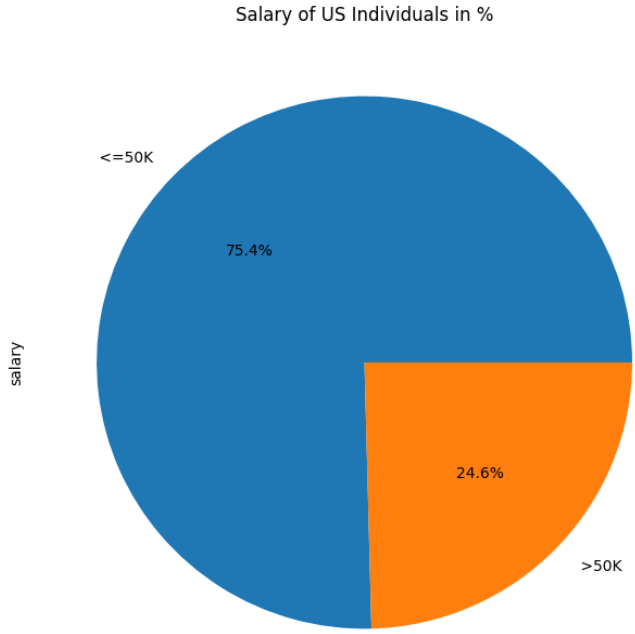

Figure 3.9: Salary of US Individuals in %.

A positive correlation between salary range in both UK and US individuals was found. While observing the pie chart we notice that over two thirds of UK and US workers amounted to less than 50K yearly salary.

Data Preprocessing

After EDA examination, we identified three problems to address:

- The duplicated values.
- The unnecessary whitespace in the values of the categorical features.
- The value '?' in the aforementioned categorical columns must be replaced by another value.

Starting with the duplicate values problem, since only 24 rows were affected, we decided to remove them. This will have minimal impact on the research results. With one line of code, we drop the duplicates in the dataset. Afterwards, we confirmed that no duplicate values remained.

Secondly, the unnecessary whitespace in the categorical values. This can also be corrected with one line of code. Making use of Python's support for the functional programming paradigm, we use the *apply* method to pass in a lambda function as an argument that eliminates whitespace from the categorical values of the dataset, evidenced in Figure 3.10.

```
[ ]  df = df.apply(lambda x: x.str.strip() if x.dtype == 'object' else x)

     for column in df.select_dtypes('object'):
       print(f'The {column} has unique values: {df[column].unique()}')

     The workclass has unique values: ['State-gov' 'Self-emp-not-inc' 'Private' 'Federal-gov' 'Local-gov' '?'
      'Self-emp-inc' 'Without-pay' 'Never-worked']
     The education has unique values: ['Bachelors' 'HS-grad' '11th' 'Masters' '9th' 'Some-college' 'Assoc-acdm'
      'Assoc-voc' '7th-8th' 'Doctorate' 'Prof-school' '5th-6th' '10th'
      '1st-4th' 'Preschool' '12th']
```

Figure 3.10: The Lambda Method Used to Eliminate Whitespace.

Lastly, replacing the '?' value in the dataset. We noticed that whenever the 'Never-worked' value was present in the *workclass* field, the *occupation* field had a '?' value. Thus, we replaced those specific values with the more appropriate 'None' value. Next, we confirmed that the '?' value existed in the same rows to replace them

simultaneously. Unfortunately, the number of '?' values in the *occupation* field matched the number of ' ?' values in the *workclass* field, the number of '?' values in the *native-country* field were different. To solve this problem, we decided to label any '?' value in the *occupation* field with 'Unknown-occupation', label any '?' value in the *workclass* field with 'Unknown-work-class', and label any '?' value in the *native-country* field with 'Unknown-native-country'. This labelling would later help us in modelling the data and lead to more accurate results.

Once the EDA problems were solved, we proceeded to the data modelling. We chose a specific supervised and unsupervised machine learning algorithm. Both algorithms will be applied to the datasets to train a model to analyse data. The results of both approaches will be discussed in detail in the following chapter.

Chapter 4

Data Modelling - Supervised Analysis

Decision Trees

Supervised learning approach focuses on predicting outcomes by using labelled data. This phase consists in developing a model that can accurately classify whether an individual's salary exceeds the $50,000 threshold based on socio-economic attributes. Firstly, we identified the target variable relevant for this project, the *salary* feature. Then, encoded all categorical features to assure algorithm compatibility and prevent possible bias. Afterwards, we split the dataset into a training set and a testing set with a 70:30 ratio, considered one of the most optimal for machine learning (Nguyen et al., 2021). Due to the nature of the dataset being categorical and numerical the suited approach for this project was using Decision Trees (Charbuty and Abdulazeez, 2021).

Supervised Model Performance Metrics

The confusion matrix table describes the correct and incorrect predictions of the algorithm facing the real dataset. Figure 4.1 is the correspondent confusion matrix to our supervised modelling approach of Decision Trees, and Table 4.2 is the classification results. For interpretation purposes, *Class 0* represents salary below or equal to $50,000, while *Class 1* represents the salary above $50,000.

From the precision ratio of our algorithm, we can extract a higher tendency of correctly predicting *Class 0* (0.88), than it has at predicting *Class 1* (0.59). Associated with the low precision of *Class 1*, there is a moderate likelihood of prediction of false positives. This suggests that the algorithm is struggling to find a pattern for high income within the current features in the dataset.

Evaluating the recall ratio, the algorithm still shows a disparity between *Class 0* (0.87) and *Class 1* (0.62). The lower recall ratio in *Class 1*, may lead to predictions of

false negatives. Conversely, on *Class 0,* the high recall value indicates a high degree of completeness in capturing the adequate instances.

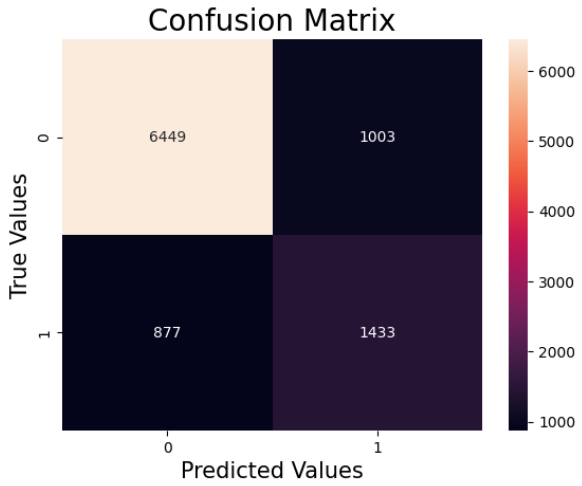

Figure 4.1: Confusion Matrix of Supervised Approach.

	precision	recall	f1-score	support
0	0.88	0.87	0.87	7452
1	0.59	0.62	0.60	2310
accuracy			0.81	9762
macro avg	0.73	0.74	0.74	9762
weighted avg	0.81	0.81	0.81	9762

Figure 4.2: Classification Report.

Unsupervised Analysis

K-Means Clustering

K-Means clustering was opted as our unsupervised learning algorithm after taking some key factors into consideration. The simplicity and scalability of K-means made it computationally efficient and straightforward in handling a big and complex dataset (Bahmani et al., 2012). Figures 4.3 and 4.4 depict the scatterplot after fitting the K-Means model to the dataset and the different metrics that were used to evaluate the accuracy of the model, respectively.

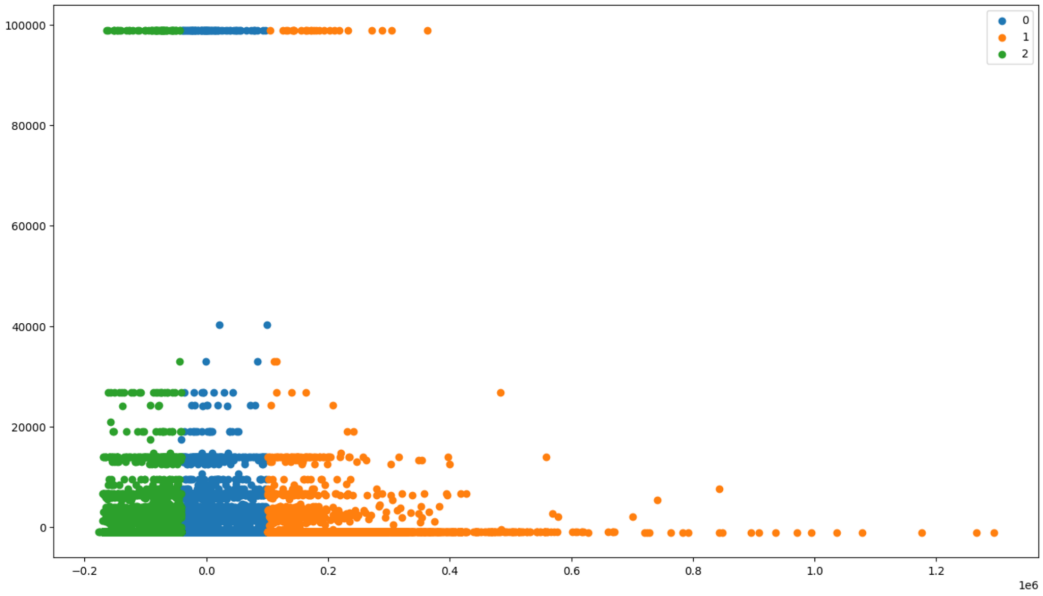

Figure 4.3: K-Means Clustering of Unsupervised Approach.

```
Silhouette Score: 0.5524327141124625
CH Score: 56274.04137876429
Completeness Score: 2.079919514936429e-05
Homogeneity Score: 3.768759416506456e-05
```

Figure 4.4: Cluster Evaluation Metrics.

Unsupervised Model Performance Metrics

Silhouette Score

Silhouette Score is used to measure how well-separated the clusters are. A higher score of 0.55 indicated that the clusters were reasonably well separated and distinct from each other.

Calinski-Harabasz Score (CH Score)

The function of CH score is to measure the quality of a cluster where a higher CH score shows that the clusters are compact and well separated. By having a score of 56,274, it further justified the existence of a strong clustering.

Completeness Score

Completeness Score is useful in measuring whether all the data points which belong to the same true class are truly assigned to the same cluster. Getting an extremely low value which was close to zero implied that the clusters did not nicely represent the ‘salary’ feature.

Homogeneity Score

Homogeneity Score on the other hand is used to measure whether all data points in a cluster belong to the same true class. A very low value calculated indicated that the clusters are not highly homogeneous with respect to the ‘salary’ feature.

Using the scatter plot and the evaluation metrics, we could say that the K-means clustering algorithm performed reasonably well in separating the data into distinct clusters. This could be proven through the grouping and separation of the data points in the scatter plot. However, the low value in both Completeness and Homogeneity Score suggested that there may still be some overlap between clusters especially with respect to the ‘salary’ feature.

Conclusion

To conclude, the aim of our study was to discover the factors leading to high salaries via the analysis of a dataset on salary predictions. The supervised model had success in predicting *Class 0*, but struggled with *Class 1*, suggesting challenges in predicting high salaries resulting from insufficient representation of high salary data within the dataset and the relevance of the features.

Moreover, K-means clustering explored potential patterns and clusters in the data and subsequently provided valuable insights. Yet, it highlighted the existence of overlap between clusters in the 'salary' feature. Besides, the scatter plot generated by K-means clustering allowed us to visualise how data points were distributed across different clusters. Contrarily, as data points are assigned to particular clusters, K-means enabled interpretable results and this could subsequently identify meaningful segments within data.

This project showed the importance of considering various factors beyond the scope of this dataset and the complexities of predicting high salaries. Primary data collection could be employed and additional years of census data could be explored in the future to enhance the understanding of high income prediction factors. Despite the limitations mentioned, this study served as a valuable starting point for identifying data patterns and segments within the dataset.

Reflections

Throughout the analysis we realised that our projects' goal was optimistic under the project circumstances. Concerning the supervised approach, one issue was insufficient data representation of salary above 50k or insufficient data. Secondly, achieving a high salary may not be so easily predicted by the fourteen features present in our database, nonetheless it may be linked with other features outside the reach of our dataset.

K-means was able to explore potential clusters, underlying patterns and groupings within the dataset which consists of both numerical and categorical data. Nonetheless, there were limitations associated with K-means where the assumption of K-means that clusters are spherical and of equal size may not be perfect fit for a complex datasets and it can be challenging to determine the number of clusters in advance. It may be worthwhile to explore alternative clustering methods to improve the interpretability and separation of clusters.

Environment and Packages

Language: Python 3.11.5

Jupyter: 6.5.4

Packages used:

- numpy
- pandas
- matplotlib
- matplotlib.pyplot
- seaborn
- warnings
- sklearn.model_selection
- sklearn.tree
- sklearn.metrics
- sklearn.cluster

References

Alonso-Almeida, M.D.M. and Llach, J., 2019. Socially responsible companies: A re they the best workplace for millennials? A cross-national analysis. *Corporate Social Responsibility and Environmental Management*, *26*(1), pp.238-247.

Asaari, M.H.A.H., Desa, N.M. and Subramaniam, L., 2019. Influence of salary, promotion, and recognition toward work motivation among government trade agency employees. *International Journal of Business and Management*, *14*(4), pp.48-59.

Bahmani, B., Moseley, B., Vattani, A., Kumar, R. and Vassilvitskii, S., 2012. Scalable k-means++. *arXiv preprint arXiv:1203.6402*.

Charbuty, B. and Abdulazeez, A., 2021. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), pp.20-28.

Didit, D.D. and Nikmah, N.R.S., 2020. The role of remuneration contribution and social support in organizational life to build work engagement. *Journal of Islamic Economics Perspectives*, *1*(2), pp.20-32.

Nguyen, Q.H., Ly, H.B., Ho, L.S., Al-Ansari, N., Le, H.V., Tran, V.Q., Prakash, I. and Pham, B.T., 2021. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, *2021*, pp.1-15.

Ozili, P.K. and Arun, T., 2023. Spillover of COVID-19: impact on the Global Economy. In *Managing inflation and supply chain disruptions in the global economy* (pp. 41-61). IGI Global.