COMP 1886 Coursework report
Student Name: Joseph Ghansah 9 Digit ID number:001409294

Task 1: Identify a scenario and perform a literature study.

In this first task, this essay will focus on job recruitment patterns for gig workers in the UK. This literature study will first give a brief insight into the gig work and mention the role of AI and machine learning algorithms in the gig work in the UK. I will then end by giving a summary of the information and insights gathered throughout the literature study.

A gig worker, according to the Oxford dictionary can be defined as, “a person who does temporary or freelance work, especially and independent contractor engaged on an informal or on-demand basis”. From the definition, one can infer that gig work is any work which is a freelance, requested only on demand and is informal. Examples of gig work which is increasingly growing are delivery services (like Deliveroo), ride sharing (uber and bolt), freelance writing and tutoring and some home maintenance services. In the UK there is a growth in the expansion of the gig economy. Wood, Martindale, and Burchell (2023) highlighted how digital platforms like Uber, Deliveroo, and Upwork have transformed the UK job market, leading to a departure from traditional employment models. They note the convenience that gig workers' services offer to consumers, contributing to their increasing demand. One of the primary reasons for the appeal of gig work, according to Cockett and Willmott (2023), is the flexibility it offers, particularly among younger workers. However, as Broughton et al. (2018) point out, gig workers often face significant disadvantages, such as the lack of access to benefits like sick pay, pensions, and job security. After this brief introduction into gig work, its growing patterns and the pros and cons, I will now delve into the role that AI and machine learning is playing or has played in the gig working economy.

Although data and models made using ML models and algorithms are quite difficult to come across, it is well known that online gig services use algorithms and machine learning models to match workers to customers or consumers. There are reports however that came from these studies although details about the research is not given. Knight, B et al (2023) talks about how machine learning is harnessed to enhance productivity on gig platforms. An example will be using Predictive models to analyse worker performance and make recommendations for task allocation based on historical data. Duggan et al (2023) also explores how algorithmic technologies shape gig work environments. An example of this will be using Machine learning algorithms for demand prediction, allowing platforms to efficiently manage workforce distribution.

The insights gained from the scenario chosen are as follows;

- From the scenario, it is very evident that ML algorithms play a key role in the online gig industry.
- Some of the ML roles in this industry are; using predictive and classification models to match gig workers to customers/ consumers.
- There are questions raised about the bias and fairness metrics of these machine learning algorithms and models as to whether they generate biases or not.

With this, it can be noted that there is more to explore about the use and misuse of Machine learning models in the gig industry which is semi-supervised and volatile.

(Citation will be done at the end of documentation).

Task 2: Exploratory Data Analysis (EDA)

COMP 1886 Coursework report
Student Name: Joseph Ghansah 9 Digit ID number:001409294

This EDA task was done to gain more insight into the dataset. In doing so, it gives you an idea of the spread of the data, finding missing values and other important things one ought to know about the dataset before pre-processing. From the EDA, I was able to find out about the spread of my dataset. Finding the data type of my numeric columns bright me to an understanding that my numerical columns have some strings in them and needed to be taken care of during pre-processing. By looking at my categorical features, I was able to fish out the numerical columns which had strings in them because they appeared in the categorical aspect also, (this is pointed out in the markdown on the notebook cell). The first EDA plot I made was the graph showing missing values. Below is the graph.

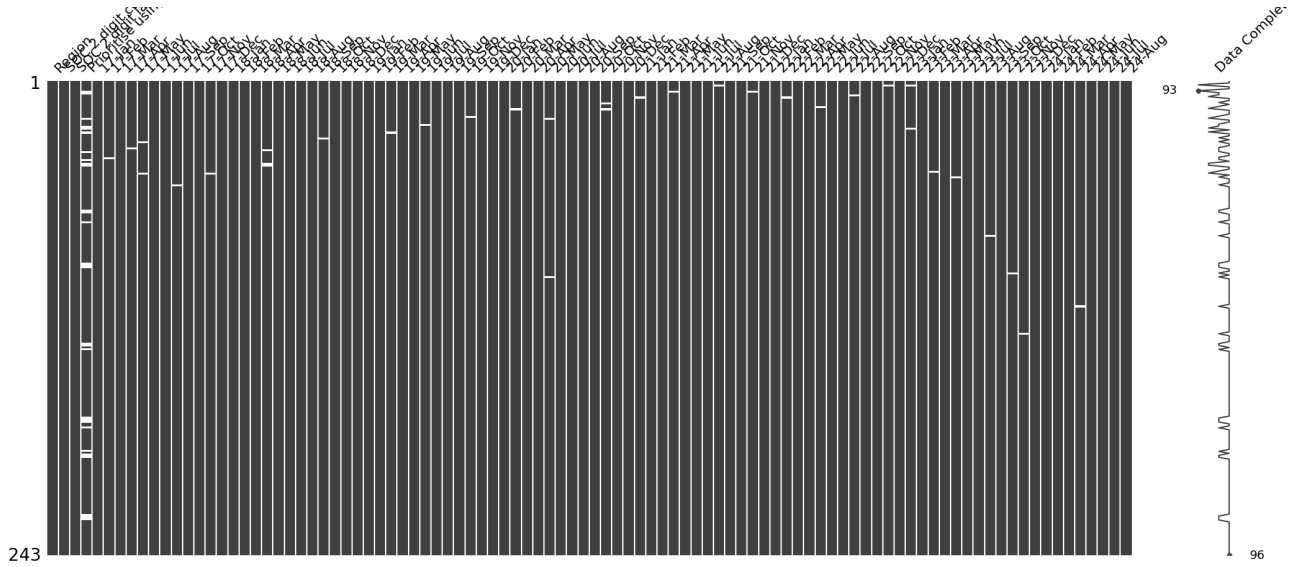

Figure 1: Graph showing the spread of missing values

From the figure above, the dataset is not complete. There are white dashed lines which indicates the presence of missing values in the dataset. However, these missing values are not so much as there is a 96% data completeness from the graph.

The next graph that this report plotted to explore the dataset was a plot that showed the overall trend of monthly job advertisement over time. From this plot, which is shown below, it can be observed that the lowest month for overall advertisement was in April 2020, with the highest being May 2022. It shows a trend which when averaged out, seems to neither be an upward or downward trend. However, there was a significant dip in advertisement between June 2018 and February 2021 as can be seen from Figure 2 below.

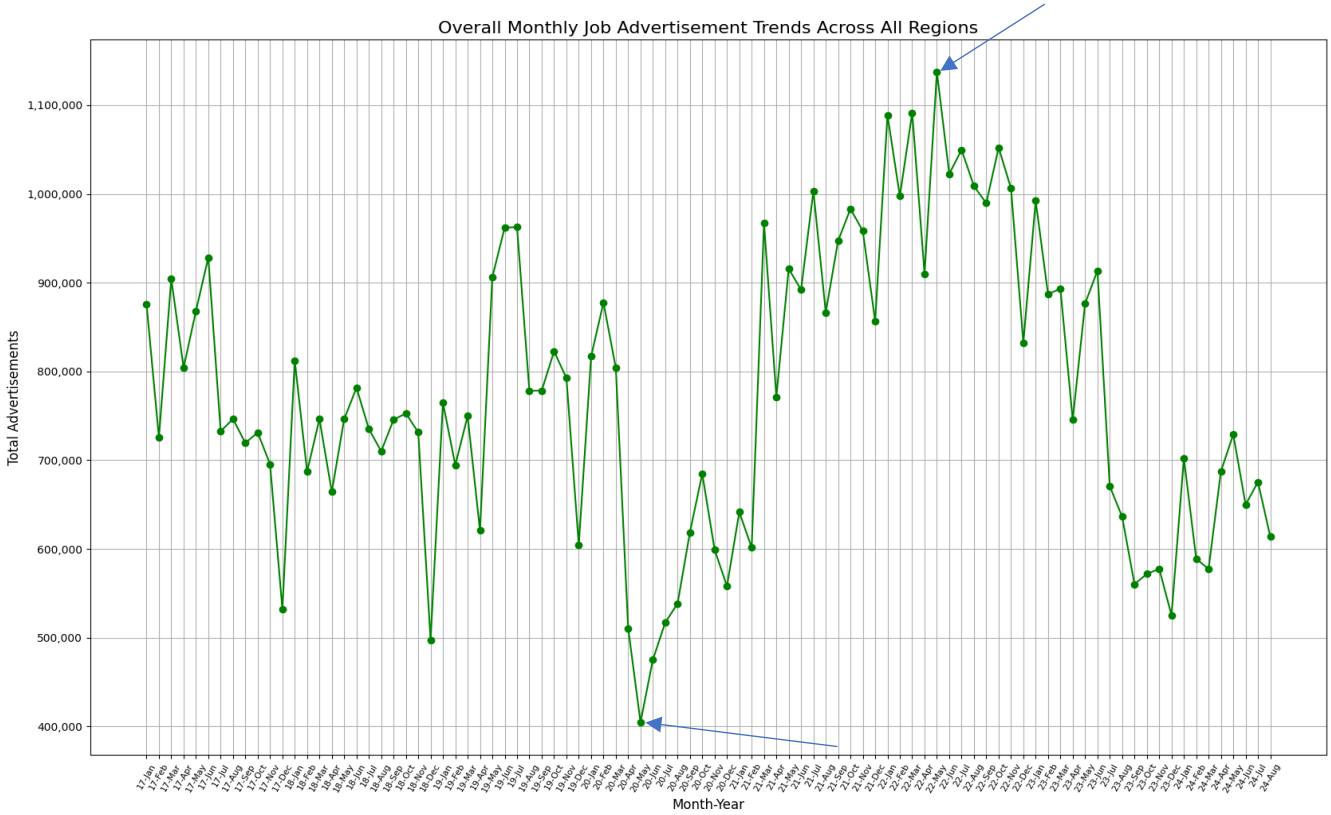

Figure 2: Overall monthly Job advertisement trend for all regions

The arrows in Figure 3 above point out the highest month for advertisement and the lowest month. The next graph is the yearly advertisement by SOC label which is in Figure 4 below.

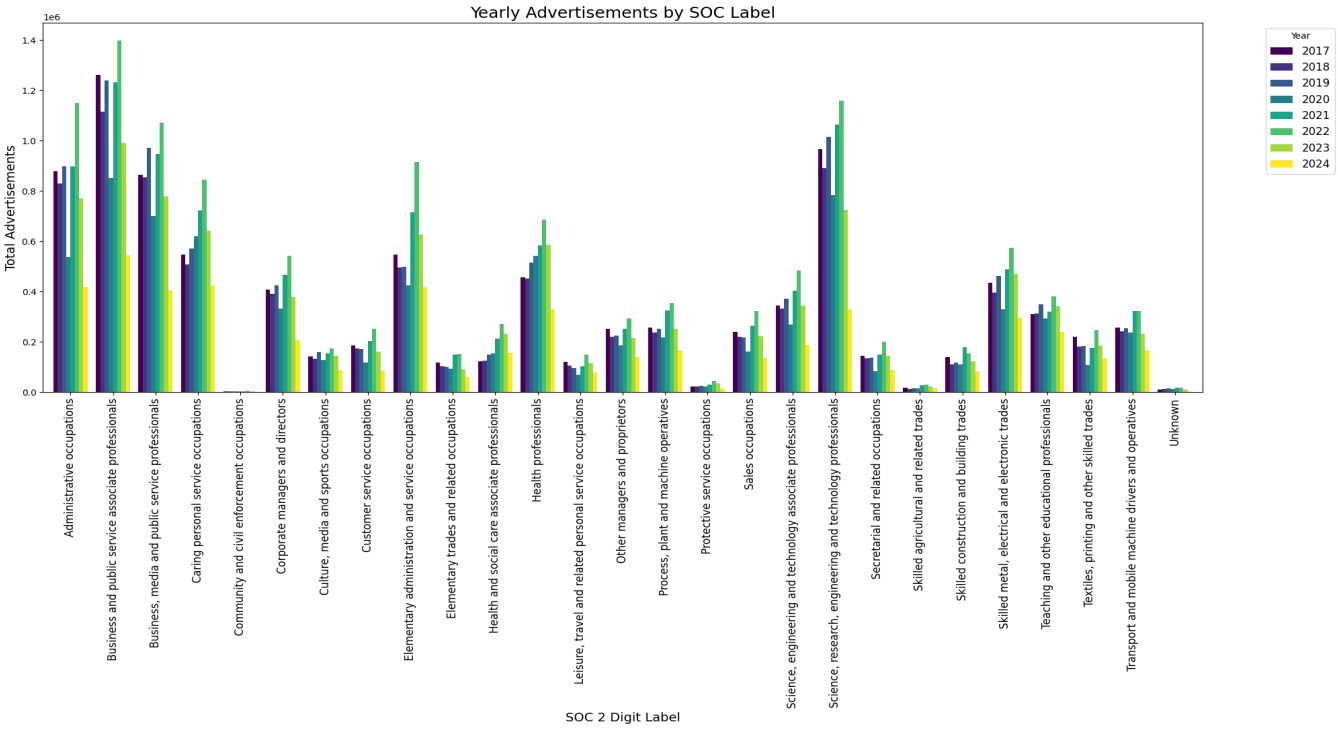

Figure 3: Yearly Advertisement by SOC label

From Figure 3 above, it can be observed that the SOC label with the highest advertisement level is the business and public service associate professionals. The lowest of them all is the community and civil enforcement occupations. It also shows by the year, how these SOC labels fared against each other. The next EDA visualization included in this report is one that looks at the total number of advertisements by region. It can be seen in Figure 4 below.

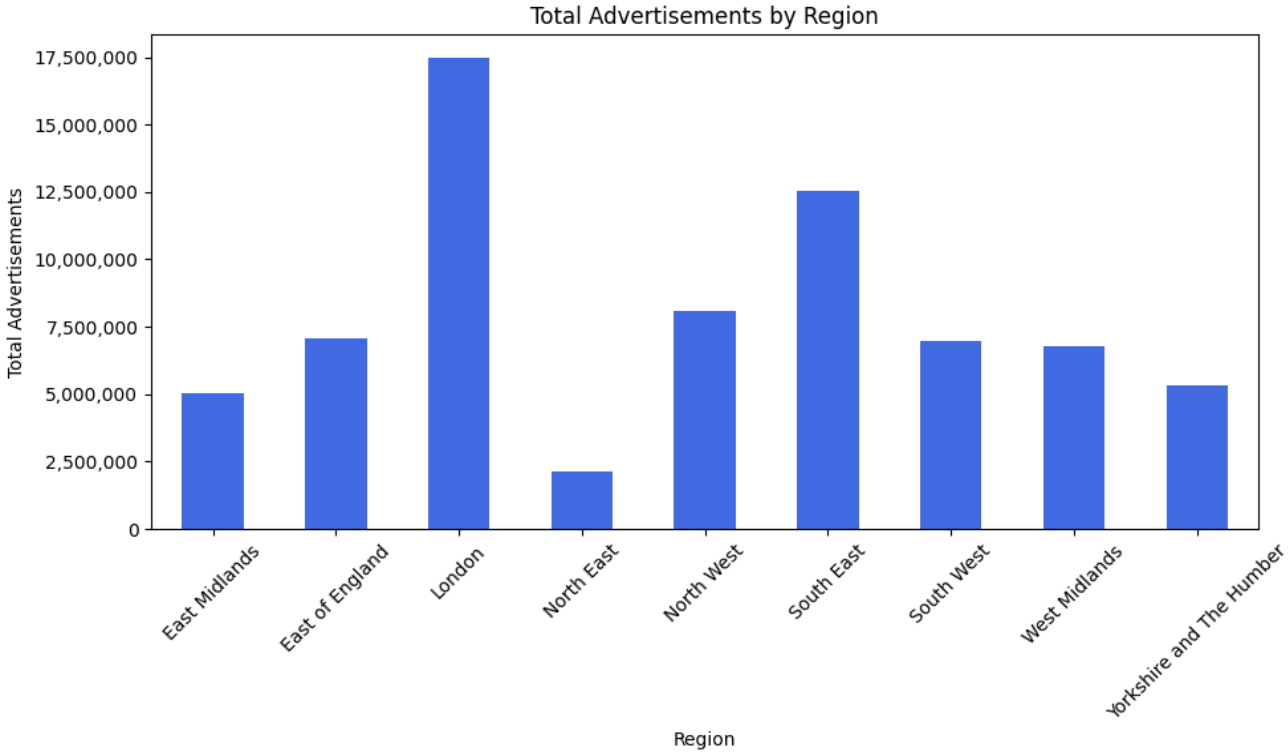

Figure 4: A bar chart showing the total advertisement by Region

From Figure 4 above, it can be observed that most of the advertisements were made in London, followed by the South East. The region with the lowest number of advertisements was the North East. The last plot for EDA in this report will look at the spread of data. Most of the months had a skewness which was above 1 which means that the data over the months was not spread evenly. The data is extremely skewed positively, most values are clustered towards the lower end of the distribution. The kurtosis of almost of the months were also above 3. This means that it is a positively skewed distribution, where the tail on the right side is longer, indicating that the data is concentrated towards the lower end of the distribution with a few larger values pulling the tail out to the right. The boxplot showing the distribution of values across SOC labels also showed the same which can be seen below.

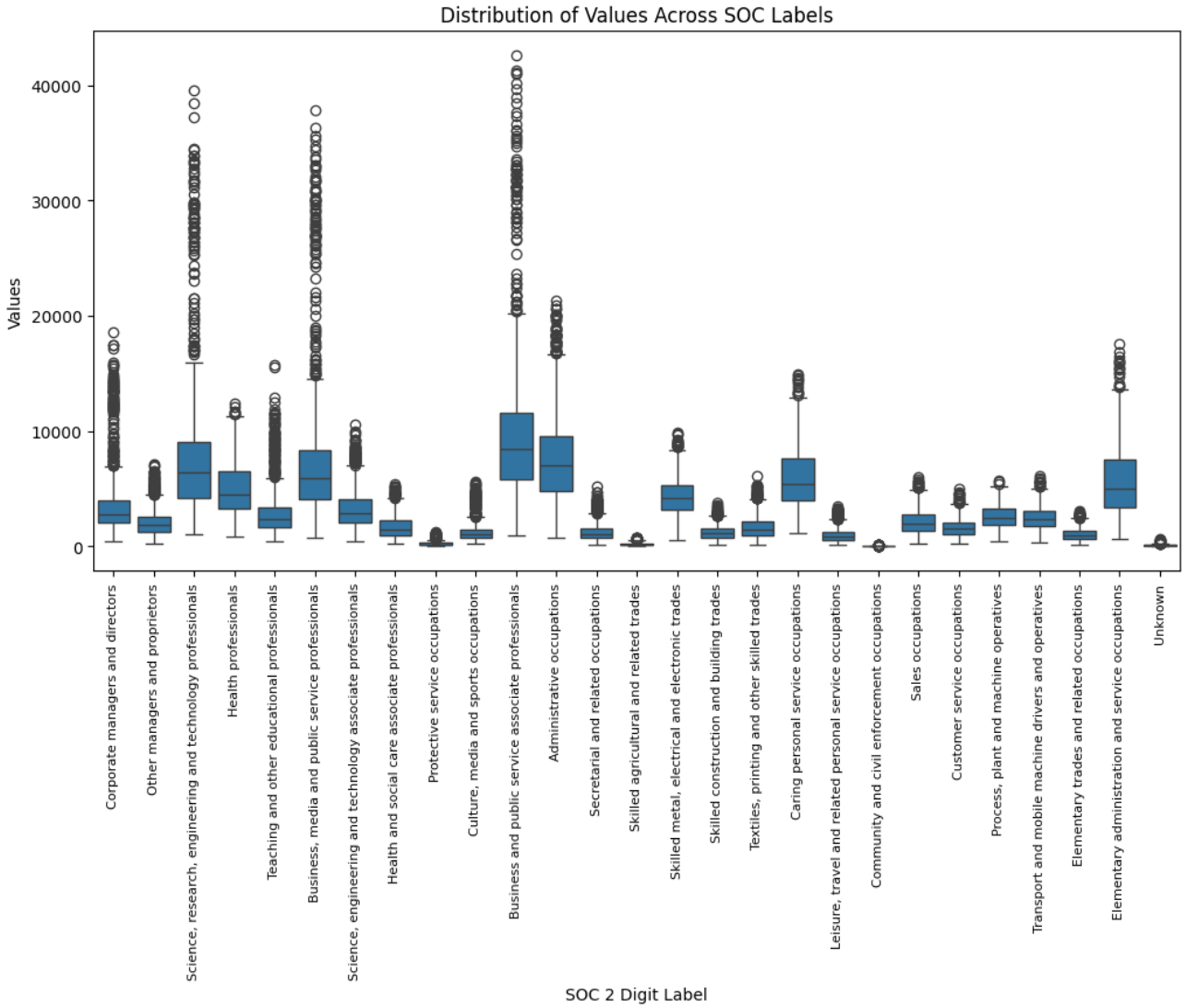

Figure 5: A box and whisker plot showing distribution across SOC labels and outliers.

From this plot, all the SOC labels have outliers in the dataset which points out that during data pre-processing, measures need to be taken to take care of these outliers that have been shown.

From the EDA above, this report has found that the following pre-processing techniques need to be employed;

- There are some features which have categorical data mixed with numerical data which needs to be cleaned to prepare the dataset for ML models.
- The dataset has outliers which need to be taken care of with an appropriate technique.
- All regions and SOC labels are well taken care of, which makes for a good dataset.

Task 3: Data Pre-processing

In this task, I did some data pre-processing to make the dataset ready for the machine learning model. The following are the pre-processing tasks that was done on the dataset.

1. The first pre-processing task was to make sure all numerical columns were only numerical and did not have categorical data. From EDA it was found that some numerical data contained strings '[x]'. The report assumes that the '[x]' signifies that there were no

advertisements in that month and hence will replace all the '[x]' in the numerical columns with a zero.

2. The second pre-processing task involved filling in the missing values. For this step, I replaced the empty data points in each numeric column with the median of that column (i.e., the specific numeric feature). Using the median helps preserve the dataset's integrity by minimizing further skewness and effectively handling outliers.
3. The third pre-processing technique this report does is fill in missing values for categorical data. Column D which is the **Prioritise using SOC 2 as high or low** has missing data and I encoded them by filling in missing datapoints with 'unknown'.
4. The next preprocessing task was to remove the B and C columns and encoding the categorical values. I used the one-hot encoding.
5. The next thing I did was normalization of the data. I used the robust normalization technique to do the normalization. From literature review, using the robust scaler helps handle outliers very well in a model without removing them. I used this scaler because I had some fear that removing outliers will lead to missing out on some of the data. Literature review suggested that using of the robust scaler or log transformation since both of them preserve the integrity of the dataset even though there are outliers. Since the dataset is a tracking of trend across time, it is good that outliers are kept, these outliers may be a trend rather than noise.

Task 4: Supervised ML algorithm: Prediction

In this part of the task, I performed a prediction with the random forest regressor. The choice of using the random forest regressor stemmed out of the fact that the initial exploration of the dataset did not prove to show that there was any linear relationship between the features (see notebook file). This necessitated the choice of using the random forest regressor. From literature review, it was noticed that tree-based methods like the random forest regressors handle complex interactions and non-linearity very well. It also is said to be less sensitive to outliers, which is evident that the dataset being worked on has, this made the random forest a good choice for the model. This report will now point out the way of proceeding using the random forest regression model.

The choice of splitting the dataset was to use the train-validate-test to allow for hyperparameter tuning. The dataset was initially split in a 60% - 40% way. The 60% was to train the model while the 40% was divided equally to validate and test the model. The following were the results based on the following parameters:

1. 50 trees: Validation RMSE: 0.2254, Test RMSE: 0.0993
2. 100 trees: Validation RMSE: 0.2132, Test RMSE: 0.0960
3. 200 trees: Validation RMSE: 0.2102, Test RMSE: 0.0964

From the results above, the best fit for the model is when it has 100 trees. It is the best compromise ensuring that the model does well when it comes to the test data while making sure it also is not underfit. The next task was to plot the Actual values Vs the predicted values. Below is an image of the plot.

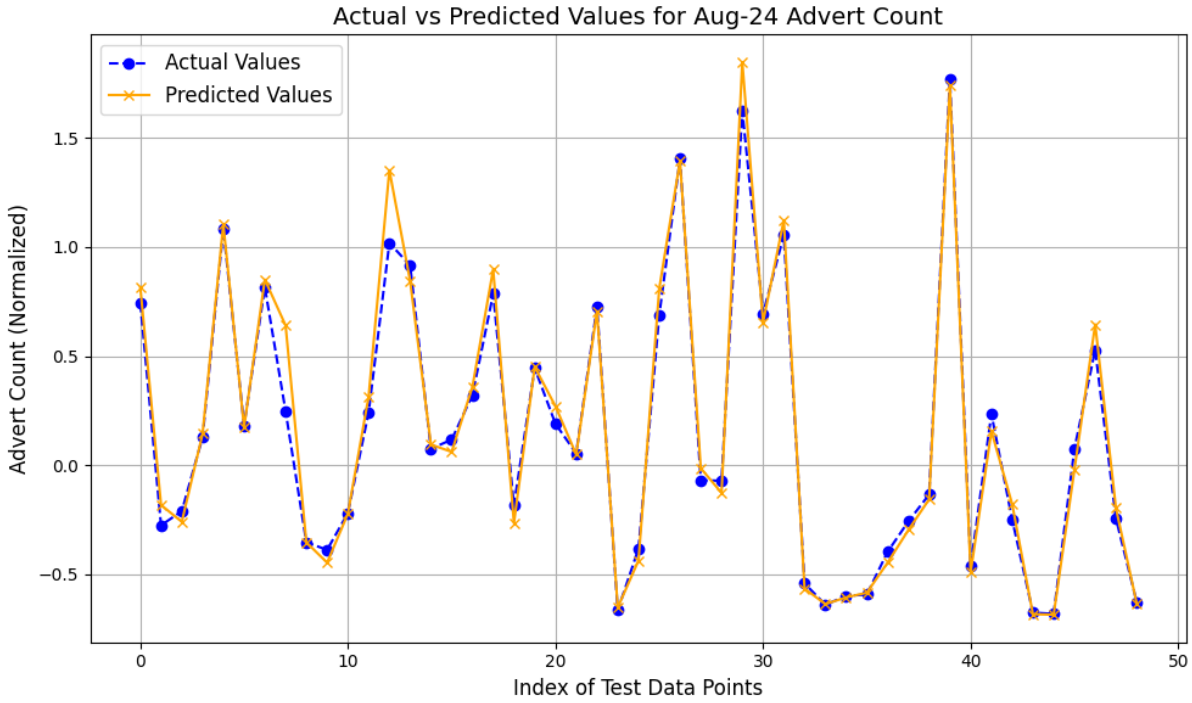

Figure 6: A plot of Actual values and Predicted Values for Aug-24

From the figure above, it can be seen that the actual values are not too far from the predicted values. This indicates that the model works quite well. Another scatter plot shows the actual values and predicted values on a scatter plot with the perfect prediction line is below.

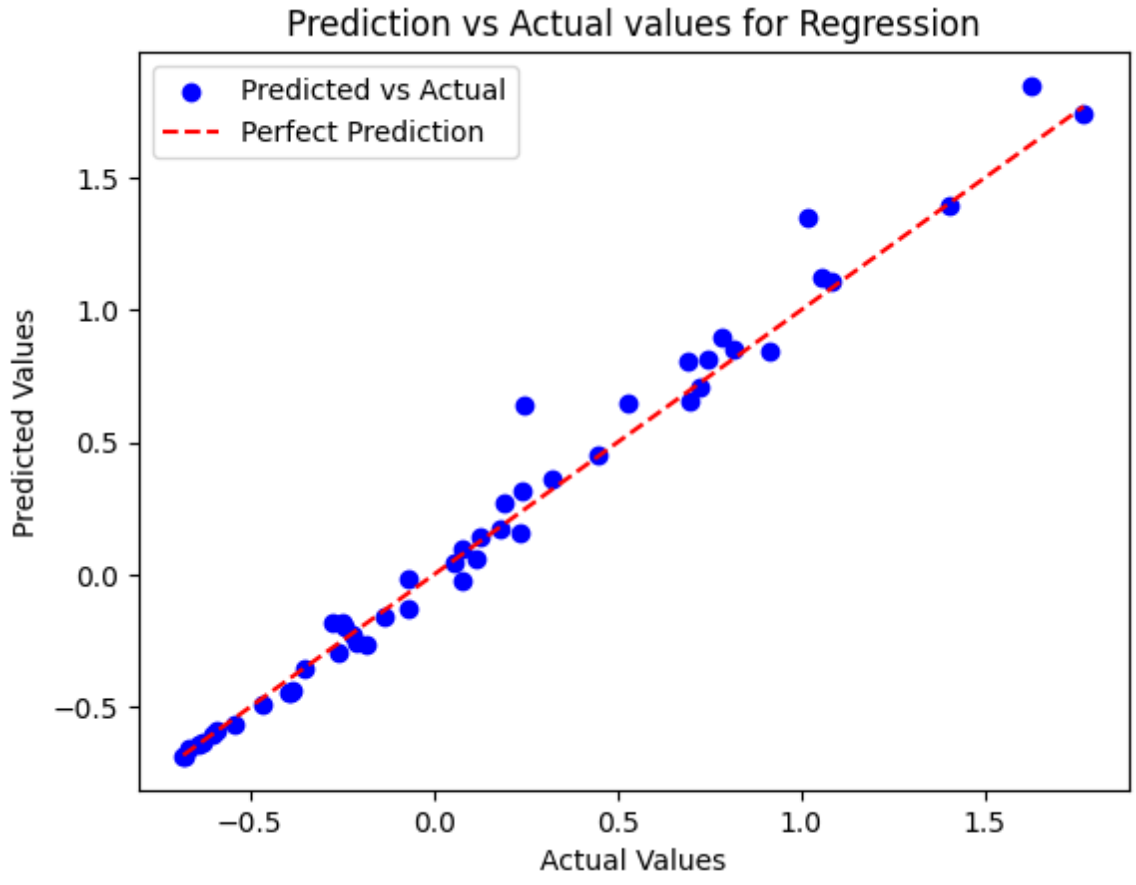

Figure 7: A scatterplot graph showing the actual and predicted values with the perfect prediction line

From this scatterplot, it can be shown that the predicted are generally in the same trend as the actual values which is a good thing in general. There are the following however that point to the fact that the model could have done better. With a training R^2 of 0.9957, it means that the model fits the training data extremely well. With a validation R^2 of 0.9307, it suggests some level of generalization also. The test R^2 of 0.9767 indicates also that the model performs well on new and unseen data also which is a good thing. The final report (in the notebook), the RSME on the test set yielded a 0.0960. This indicates the average error magnitude and this is very low. This shows that the model's predictions are close to the actual values. The MAE on the test data which is the absolute mean error was 0.0596 which suggest a minimal prediction deviation. This model will work well with some more parameter tuning. This essay will now move on to the classification model and point out the steps used.

Task 5: Supervised ML algorithm: Classification

The next model that was worked on was the classification model. The choice of the SVM model was because from literature review, the SVM model is good at catching complex relations within the dataset. The flaw of using the SVM is that it works on larger datasets. The dataset being worked on for this coursework is slightly smaller. The process of running the classification model is as follows. Firstly, I imported the necessary libraries for the SVM model. I then dropped the B and C columns. For a better classification result, I dropped the empty row in the target predicting column. Selecting the target and features, I normalized the numerical columns using the robust scaler. This was used to preserved outliers whiles smoothing them out. I used the label encoder to encode the categorical target variable. After these steps, the class distribution was 104 to 110 which shows a fair balance between the two classes. Doing different iterations of splits, I noticed that the model works better if the split is 75% train and a 25% test. Initializing the classifier with a balanced class weight, I run the model. The model has an accuracy of 0.5370 which means that the model does not perform well. This was envisaged as the correlation matrix between the features and the target variable yielded very low correlation values. The confusion matrix was performed on the test set, and the result is as follows;

Confusion Matrix:
[[6 21]
 [4 23]].

- The model was able to truly predict 6 classes of SOC high.
- The model incorrectly predicted 21 classes of SOC high as SOC low.
- The model incorrectly predicted 4 classes SOC low as SOC high.
- The model was able to correctly predict 23 classes of SOC low.

Below is the classification report;

```
Test Accuracy: 0.5370
Classification Report:
              precision    recall  f1-score   support

           0       0.60      0.22      0.32        27
           1       0.52      0.85      0.65        27

    accuracy                           0.54        54
   macro avg       0.56      0.54      0.49        54
weighted avg       0.56      0.54      0.49        54
```

With the classification report above, it can be observed that for the 0 class which is the high, it was observed that from the recall rate, only 22% of the data was accurately predicted, however for class 1 which is the low, 85% was accurately predicted, therefore the model is better at predicting SOC low

values than the high values. With an accuracy of 0.53, this model is not a good model. This was expected as the EDA done showed no correlation between the features and the target selected for the SVM model to train on.

Task 6 – Conclusion

This task was a very interesting task. One of the major realisations taken from this task was that it takes more than running the code or algorithm. There needs to be a high level of intentionality in the selection of scalers, the kind of encoding you do, how you deal with missing values and a lot of things. This requires a certain level of being critical which is beyond running the code and making it functional. For task 4, one of the models that I could use would have been the neural network models. From the very beginning of the tasks (EDA) I noticed that the relationship between the features(dataset) was not linear, and the complexity was even more pronounced after I did the correlation matrix. I believe that the neural network model will be able to capture more relationships than the SVM and the random forest regressor. If tasked to do this again, one of the things I will do will be to improve the quality of the dataset. I believe that there might be some other information which when added to the dataset will help improve its quality. Another thing I will consider doing will be to run the model using different iterations and selecting the best one. The scaler I selected for the two models and the encoding type I chose were due to literature review which also suggested other labelling and encoding techniques. My work would have been more robust if I did iterations of different encoders and scalers to see how well they performed and selecting the best one. One thing I did and was happy about was the fact that I was able to play around with hyperparameter tuning, I however, due to lack of time did not explore it as deeply as I would have, nevertheless, I am glad that I did and I have a good starting point to creating models that are good and created with intentionality. One of the issues that I had during the performance of the task was the battle between having good, functional codes which are clean and tracing back my work. It can be seen from my notebook file that although the structure of my codes and selection of variables are known to me, it would be difficult for another to follow without clear and concise instructions.

References

1. Broughton, A., Gloster, R., Marvell, R.A., Green, M., Langley, J. and Martin, A. (2018) *The experiences of individuals in the gig economy*. London: Department for Business, Energy and Industrial Strategy. Available at: https://assets.publishing.service.gov.uk/media/5a7b1c2240f0b66a2fc053a3/171107_The_experiences_of_those_in_the_gig_economy.pdf(Accessed: 12 July 2024).

2. Cockett, J. and Willmott, B**.** (2023) *The gig economy: What does it really look like?* London: Chartered Institute of Personnel and Development (CIPD). Available at: https://www.cipd.org/uk/knowledge/reports/gig-economy/ (Accessed: 12 March 2025).

3. Oxford University Press (2025) 'Gig worker, n.', *Oxford English Dictionary*. doi:10.1093/OED/1470018560.

4. Wood, A.J., Martindale, N. and Burchell, B.J. (2023) *Gig rights & gig wrongs: Initial findings from the Gig Rights Project*. Bristol: University of Bristol. Available at: https://www.bristol.ac.uk/media-library/sites/business-school/documents/Gig%20Rights%20&%20Gig%20Wrongs%20Report.pdf (Accessed: 12 March 2025).

5. Knight, B., Mitrofanov, D. and Netessine, S. (2023) 'AI-enabled Technology and Gig Workforce: The Role of Experience, Skill Level, and Task Complexity', *SSRN*. Available at: SSRN.

6. Duggan, J. and Jooss, S. (2023) 'Gig Work, Algorithmic Technologies, and the Uncertain Future of Work', in Lynn, T. et al. (eds.) *The Future of Work*. Palgrave Studies in Digital Business & Enabling Technologies. Springer. Available at: SpringerLink.

Declaration of AI Use

I have used AI (Any online AI tool such as ChatGPT) while undertaking my assignment in the following ways:

• To research and understand the questions in the coursework – Yes
• To research and correct codes and scripts – Yes
• To create an outline and get more ideas of the topics (Task 1)– Yes
• To explain the concepts – No
• To support and revise my use of language – Yes
• To summarise the following (list your reference) articles/resources: No
• For the tasks, I used the assistance of AI for debugging my codes to make sure I got the desired results I wanted to achieve.