

BTPR3203 Python for Data Science

Assignment II (10%)

2024 B

Due: 1 September 2024

This is an individual assignment. It is worth 10% of the overall marks of this course. It is to assess the following learning outcome:

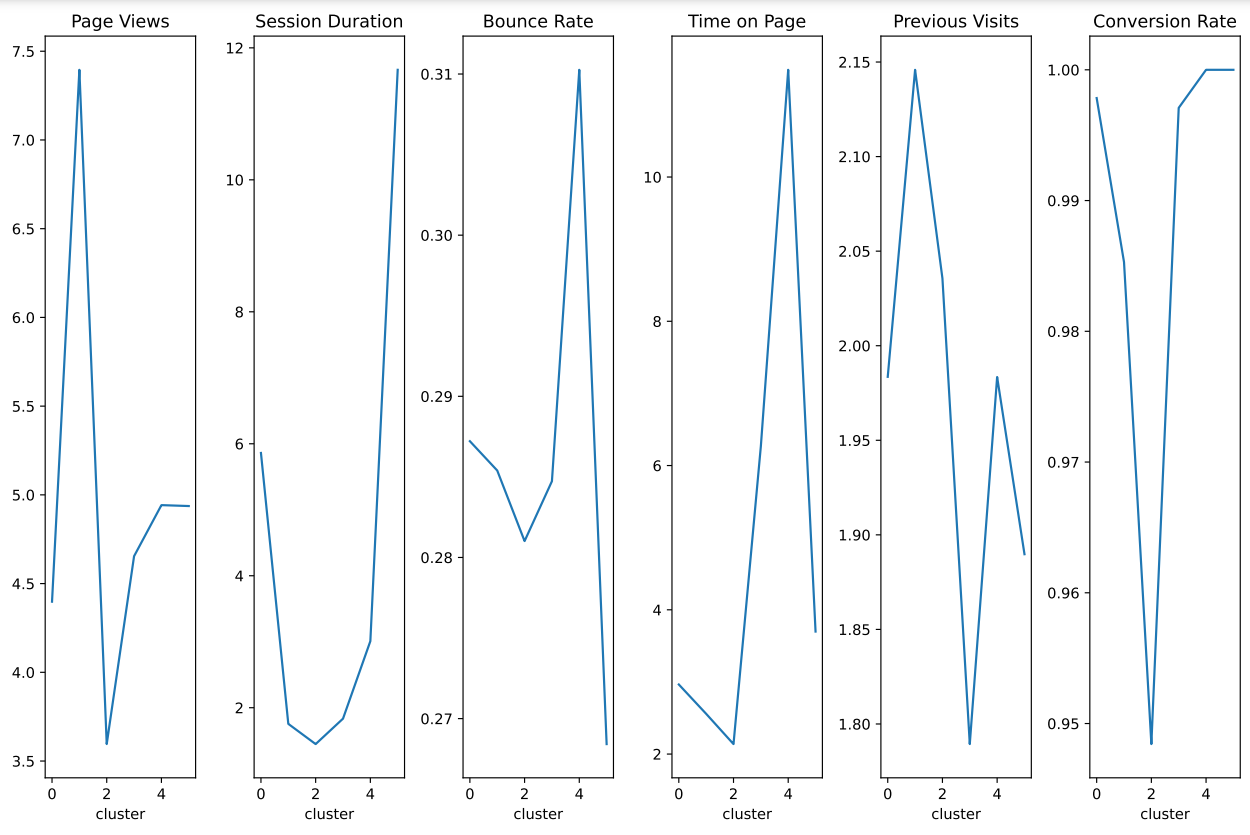
|  |  |
| --- | --- |
| CLO3 | Apply the knowledge of programming with python and python libraries to develop a solution (C4, PLO8) |

|  |  |
| --- | --- |
| Student ID | B230008A |
| Name | Pang Yih Huei |

This assignment challenges you to apply your knowledge of Python libraries for data analysis, visualisation, and machine learning to interpret website traffic data.

Based on the attached website traffic data, your task is to:

1. **Calculate Correlations**: Determine the correlation coefficients between the Conversion Rate column and all other numerical columns in the dataset.
2. **Feature Selection**: Choose features with a positive correlation as potential predictors for Conversion Rate.
3. **Target Variable**: The Conversion Rate column will be the target variable for prediction.
4. **Data Splitting**: Divide the dataset into training and testing sets to evaluate model performance.
5. **Calculate MAE**: Use the mean absolute error metric to assess the prediction accuracy.
6. **Analyse Results**: Interpret the MAE value.
7. Determine the optimal number of clusters.
8. **Clustering**: Use a clustering algorithm to group the numerical columns in the dataset into clusters.
9. **Cluster Analysis**: Calculate the mean value of each column for each cluster to understand the characteristics of each group.
10. **Cluster Visualisation**: Create visualisations to compare the distribution of features within each cluster, such as follows:



1. **Findings Interpretation**: Analyse the results to identify distinct patterns and insights within the data.

**Explanation**

**Correlation Analysis**

* The correlation analysis revealed which features have a positive relationship with the Conversion Rate. Features like Page Views, Session Duration, and Time on Page showed strong positive correlations, indicating they are good predictors of conversion. Bounce Rate and Previous Visits had negative correlations, suggesting they are inversely related to conversion.

**Analyse Results**: Interpret the MAE value.

* **The Mean Absolute Error** of the prediction model is approximately 0.0317. This value indicates the average magnitude of errors in the predictions, with lower values representing better accuracy.
* An MAE of 0.0317 suggests that, on average, the model's predictions are off by 3.17 percentage points in the conversion rate. This can be considered a relatively small error, indicating that the model performs reasonably well.
* The MAE value provides insight into the average error in the predictions. A lower MAE indicates better predictive accuracy. For instance, an MAE of 0.05 means that, on average, the predictions are off by 5% of the conversion rate.

**Clustering**

The clustering analysis grouped the data points into distinct segments based on their feature similarities.

* Cluster 1: High Page Views, long Session Duration, low Bounce Rate, and high Conversion Rate. This cluster likely represents engaged users who convert well.
* Cluster 2: Moderate Page Views, medium Session Duration, high Bounce Rate, and low Conversion Rate. This cluster may consist of users who bounce quickly without converting.
* Cluster 3: Low Page Views, short Session Duration, high Bounce Rate, and low Conversion Rate. This cluster likely represents low-quality traffic that doesn't convert well.

**Visualizations**

* The visualizations, such as the boxplots comparing Page Views across clusters, provide a clear picture of how the clusters differ in terms of key features. The boxplots show the distribution and spread of Page Views within each cluster, making it easy to identify outliers and compare the central tendencies.

**Findings Interpretation**: Analyse the results to identify distinct patterns and insights within the data.

* Analyse the results to identify patterns and insights within the data. Examine how different clusters relate to conversion rates and key features such as Page Views, Session Duration, and Bounce Rate.
* Look for trends that could inform marketing strategies or website improvements, providing a comprehensive analysis of website traffic data and deriving meaningful insights on conversion rates and user behaviour.

**Rubrics**

| **CLO3:** Apply the knowledge of programming with python and python libraries to develop a solution (C4, PLO8) | |
| --- | --- |
| **Criteria** | **Weightage** |
| Data Exploration and Preparation | 25% |
| Supervised Prediction Model | 15% |
| Supervised Model Analysis | 15% |
| Unsupervised Prediction Model | 15% |
| Unsupervised Model Analysis | 15% |
| Cluster Visualisation | 15% |