My Name: Manuvardhan Madala

Student ID: 23122728

University: University of Hertfortshire:

Course Name : Data Science

August 2024

Contents

1 REPORT 4

2 APPENDIX 9

List of Figures

1.1 Elbow Method of K Means Clustering 5
1.2 K Means Clustering(k=3) 5
1.3 Distribution of Target Variable 6
1.4 Scatter Plot: Air Temperature vs Rotational Speed 6

Chapter 1

REPORT

Clustering and line fitting are some of the most important techniques in data analysis, which help in the extraction of patterns, trends, and relationships within datasets. Clustering, especially methods such as K-Means, is an unsupervised learning technique that groups data points based on similarities, thus enabling analysts to find hidden structures or categories in the data [Saxena et al., 2017]. That is of particular help in areas such as market segmentation, anomaly detection, or customer profiling, when one would look to discover patterns with no preconceived labels. Among other methods, K-Means clustering enjoys popularity for the simplicity and efficiency of its use, especially on big datasets [Rai et al., 2010]. Line fitting, on the other hand, is the method of supervised learning that deals with regression analysis, modeling the general relationship between two variables. By fitting a straight line-linear regression-analysts can project outcomes and forecasts or comprehend how changes in independent variables will affect the dependent variable. This kind of technique finds broad applications across economics, finance, or healthcare due to the serious need to understand and predict trends in these areas [Andrade and Estévez-Pérez, 2014].

Together, clustering and line fitting provide powerful tools for understanding data from different angles: clustering reveals natural groupings within the data, while line fitting provides a clear, quantitative relationship between variables.

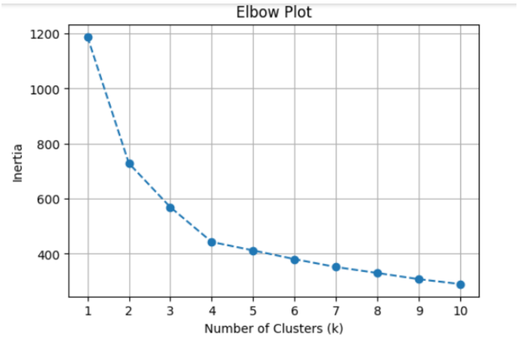

Figure 1.1: Elbow Method of K Means Clustering

Above diagram 1.1 shows Elbow Plot is used to determine the value of k for the KMeans clustering. This plots the inertia (sum of squared distances of samples to closest cluster center) against the number of clusters. A typical inertia-against-clusters plot normally shows a rapid decrease in inertia until a certain number of clusters, beyond which the decrease rate drastically flattens out and forms an "elbow." The location of the "elbow" usually suggests the optimal number of clusters: beyond this value, increases in the number of clusters lead to limited gains.In the given plot, the inertia drops steeply from k=1 to k=3 and then stabilizes. The "elbow" seems to happen at k=3, indicating that three clusters are likely to give the best trade-off between inertia reduction and simplicity. This interpretation is in line with statistical considerations of intra-cluster variance minimization while preventing overfitting.

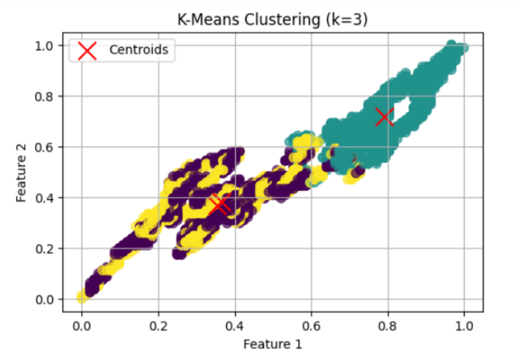

Figure 1.2: K Means Clustering(k=3)

Above diagram 1.2is scatter plot which displays the results of KMeans clustering with k=3 on a two-dimensional data set. Each point represents a data sample and is colored according to the cluster it belongs to: purple, yellow, or cyan. The red ’X’ markers show the centroids of the three clusters. It looks like a pretty nice clustering, with data points gathering around their centroids such that minimal overlap between clusters may happen. The distribution shows some sort of clear separation amongst the clusters, which supports the fact that KMeans successfully identified the distinct cluster-like groupings in this data based on the provided features. This plot supports insight into the performance of the unsupervised clustering algorithm with the number of clusters in use here (k=3) being appropriate, representing the hidden structure in the data.

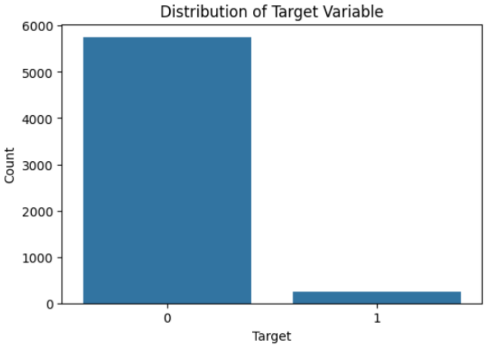

Figure 1.3: Distribution of Target Variable

Above diagram 1.3 bar graph showing the distribution of the target variable in the dataset. It represents the x-axis as a target class and the y-axis for the count of instances each class has. The plot indeed shows class imbalance with more instances of class 0 than that of class 1. Concretely, it has about 5,800 examples of class 0 and far fewer of class 1, seeming to be under 200. This can thus be viewed as a strongly imbalanced dataset in favor of class 0, and it could affect the behavior of a machine learning model since they could become biased towards the majority class. Resampling techniques or choosing appropriate evaluation metrics will play an important role in balancing this and creating robust models.

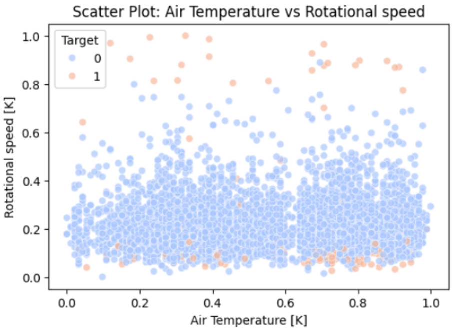

Figure 1.4: Scatter Plot: Air Temperature vs Rotational Speed

Above diagram 1.4 is scatter plot shows the relationship between air temperature (x-axis) and rotational speed (y-axis), coloring the data points according to the target variable, 0 or 1. Most of the points belong to target 0 (blue), whereas the representation for target 1 is sparse (orange). The data points are well spread over the air temperature range, indicating that there is no specific pattern or strong correlation between temperature and rotational speed. However, target 1 points appear slightly concentrated at higher rotational speeds, indicating some differentiation between the two targets within this range. Overall, this plot shows variability in the rotational speed for all air temperatures with overlapping target values.

Bibliography

[Andrade and Estévez-Pérez, 2014] Andrade, J. and Estévez-Pérez, M. (2014). Statistical comparison of the slopes of two regression lines: A tutorial. *Analytica chimica acta*, 838:1–12.

[Rai et al., 2010] Rai, P., Singh, S., et al. (2010). A survey of clustering techniques. *International Journal of Computer Applications*, 7(12):1–5.

[Saxena et al., 2017] Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., and Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, 267:664–681.

CHAPTER 2

APPENDIX

Github Link :- Github Link