UNIVERSITY OF LONDON

LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

ST3189 Machine Learning

Coursework Project

Lev Barbash

lb237@student.london.ac.uk

Raanana

2024

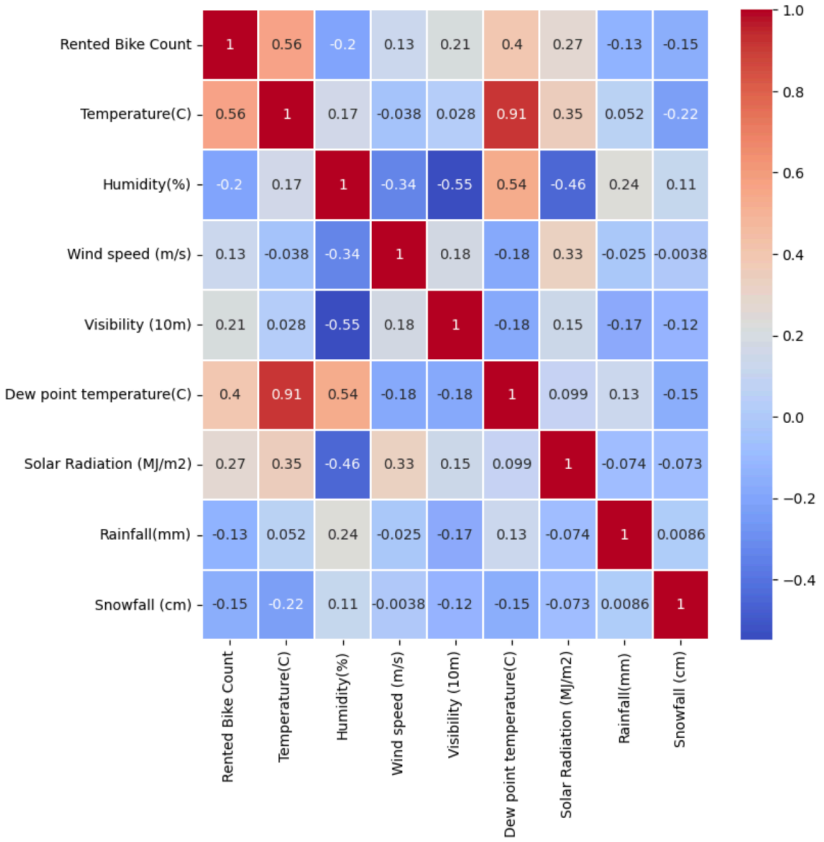
Table of contents

Regression task: Bicycle sharing demand problem

The Seoul Bike Sharing Demand dataset is used to complete the machine learning regression task. The dataset contains information about the number of shared bikes and the weather conditions during each hour of the day from December 2017 to November 2018. Also, the dataset contains some useful categorical features specifying each date (Seasons, Holidays, Functional days).

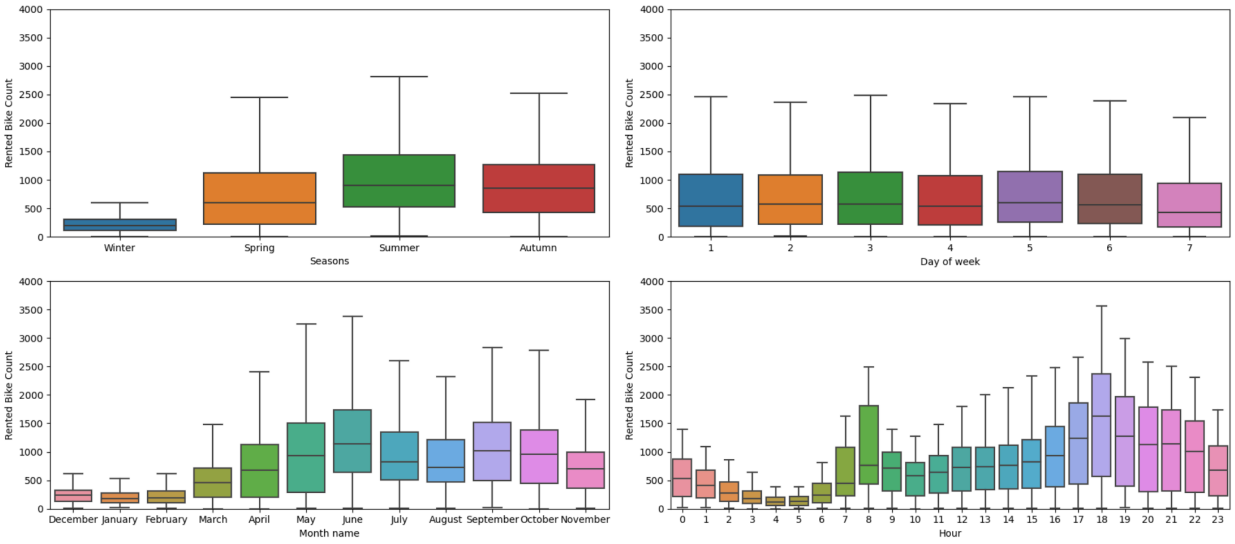
The task of the model is the prediction of shared bikes for a given hour of the day. The count of rented bikes is a target value of the dataset. The model would help to provide the city with a stable supply of bikes and lessen the waiting time for renters.

The research of the data is a valuable step before developing the predictive model. Firstly, the correlation analysis is shown below:



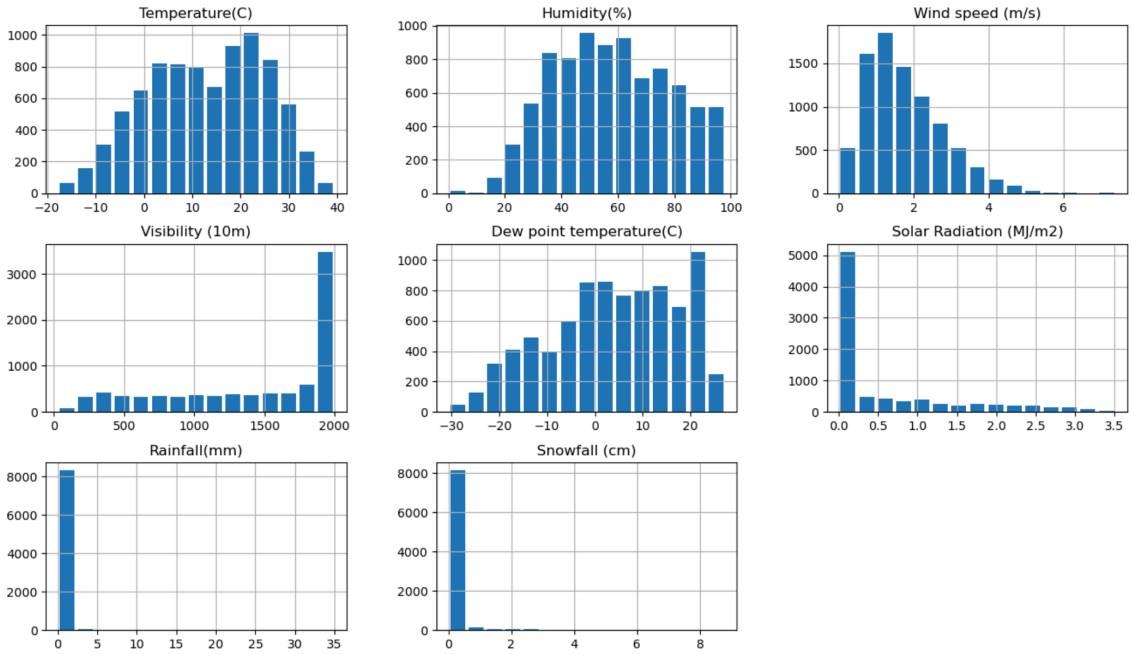
The high positive correlation between air temperature and the target value is observed. Also, the number of rented bikes has a moderately positive correlation with visibility and solar radiation and a moderately negative one with humidity. Moreover, the dew point temperature has a significantly high positive correlation with air temperature. It has to be dropped from the dataset before the linear model training process to avoid the multi-colinearity of the features. The correlation analysis points to a direct dependence between warm weather and the number of rented bikes.

The visualisation of rented bike distributions by different time points can also give valuable insights and estimate the conclusion of the correlation analysis. Thus, the box-plots graph is shown below:



The box plot is a simple method of visualising distribution. Summer has the highest distribution of the number of bikes rented while winter has the lowest one. It points to the influence of temperature on bike renting popularity. Also, research by hours has 2 periods of significant increases in distributions. Citizens rent bikes to arrive at their jobs at 8 a.m. and to return home at 6 p.m. Evening hours have stably high distributions as citizens use bikes in their free time after doing day business.

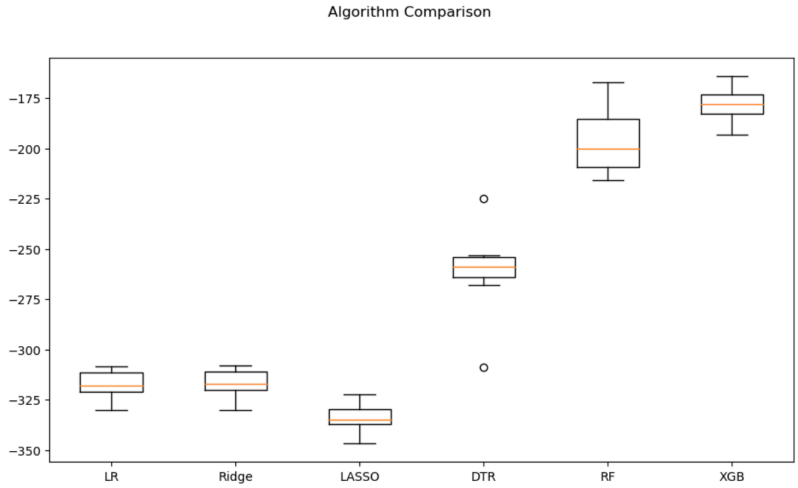
It is valuable to understand the shapes of distributions of numerical features before model training. Thus, the distributions of the numerical features are shown below:



Preprocessing of the numerical features can increase the quality of prediction. Yeo-Johnson transformation is used to make the features more normally distributed as there are negative values. Yeo and Johnson (2000) propose the following transformation:

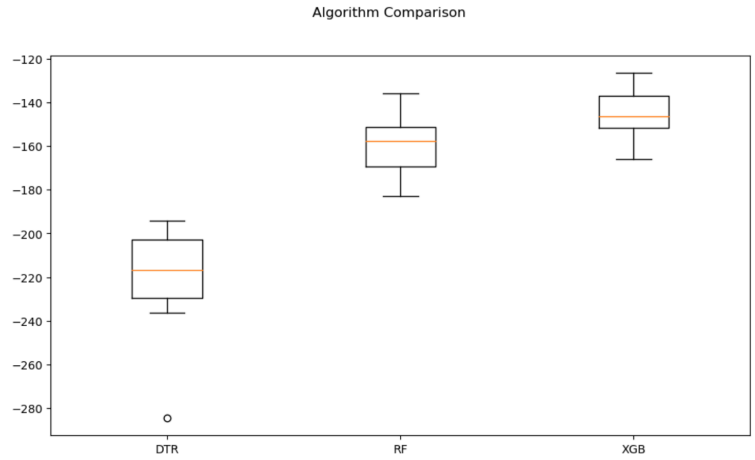
$$\psi(y,\lambda) = \begin{cases} \dfrac{(y+1)^{\lambda}-1}{\lambda} & y \geq 0 \text{ and } \lambda \neq 0, \\ \log(y+1) & y \geq 0 \text{ and } \lambda = 0, \\ -\dfrac{(-y+1)^{2-\lambda}-1}{2-\lambda} & y < 0 \text{ and } \lambda \neq 2, \\ -\log(-y+1) & y < 0, \lambda = 2. \end{cases}$$

The value of λ is chosen via maximum likelihood estimation. Also, categorical features have to be encoded for using linear models. Thus, the One-Hot-Encoding method is used. The method turns values of categorical features into features and assigns them 0 or 1 values. As the preprocessing is made, the data is ready for model training. The cross-validation method will help to choose the best model. The negative root squared error is chosen as the model quality score. Thus, the graph box plot containing distributions of scores of different models trained on 10 random splits of the data on training and test set is shown below:

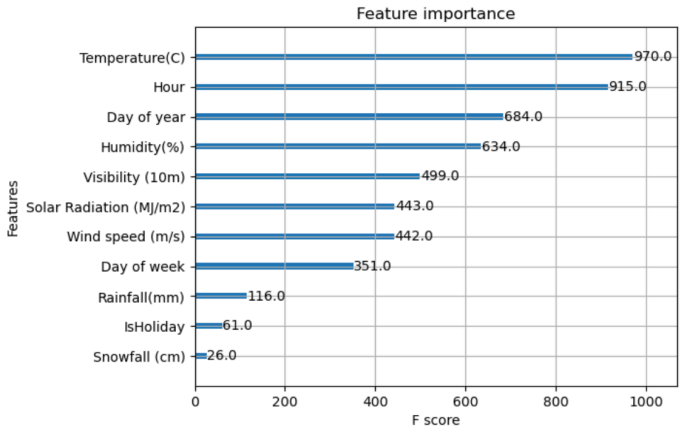


Extreme Gradient Boosting appears to be the best model for the task, the mean score of it equals *-178,53.* Random Forest regressor also shows a high result, while Decision Tree one has a moderate score distribution. Linear models (Linear Regression, Ridge, LASSO) demonstrate the worst results. Tree and ensemble of trees models seem to work significantly better than linear ones. So, the Decision Tree, Random Forest and XGB have been trained separately without numerical features transforming, as they can work well with original data. However, categorical features are encoded using Lable Encoding. The method turns values into numbers from 0 to

number_of_categories - 1. The models are trained multiple times again and the new box plots of score distributions are shown below:



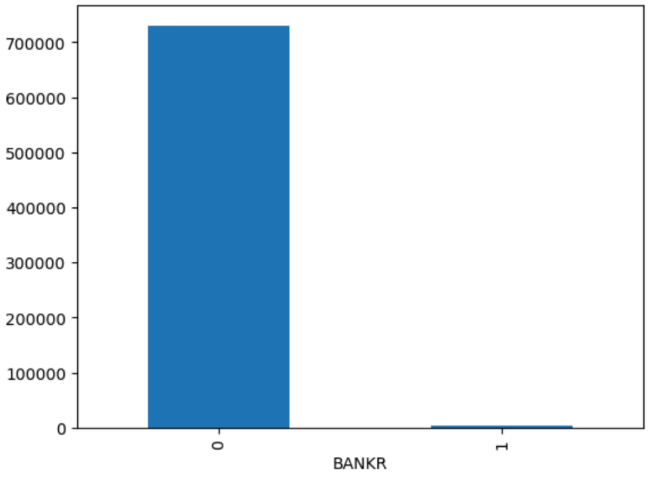
The mean score of the best model has increased slightly, now it equals *-145,39*. This is a very high result for the task of predicting the number of rented bikes for a given hour of the day. An interpretation of ensemble models is more complicated than that of linear ones. However, there is a method of estimating the importance of features. As XGB splits the data on different features during the process of building trees times a feature used for a split across all trees can be counted as a score of the importance of the feature. Thus, the graph containing the importance scores of features is shown below:



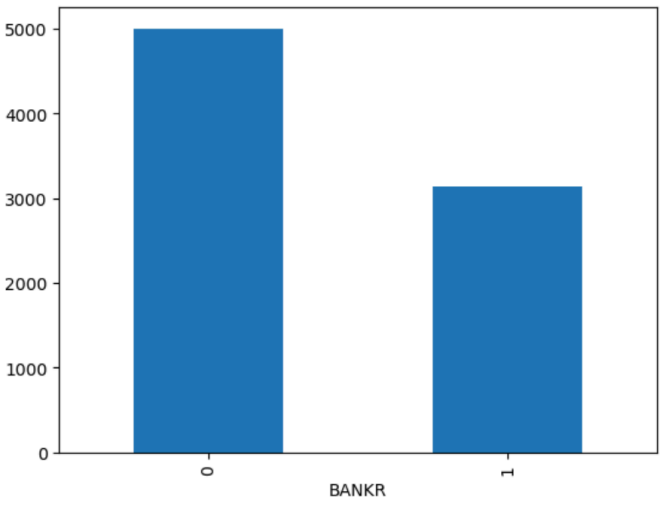
The temperature and hour as the most important features for the model prediction confirm the conclusions made before about the popularity of bike renting dependency on temperature and time of the day.

Classification task: Bankruptcy of Russian Companies prediction

The dataset containing financial information of Russian companies is chosen for the classification task completion. The values of the assets, liabilities and profits of the companies are presented for the beginning and the end of the 2017 year. The target value of the dataset is a binary feature - whether a company went bankrupt or not. The dataset is sparse (has significantly more zero values than nonzero). Moreover, the target value distribution is unbalanced as shown below:

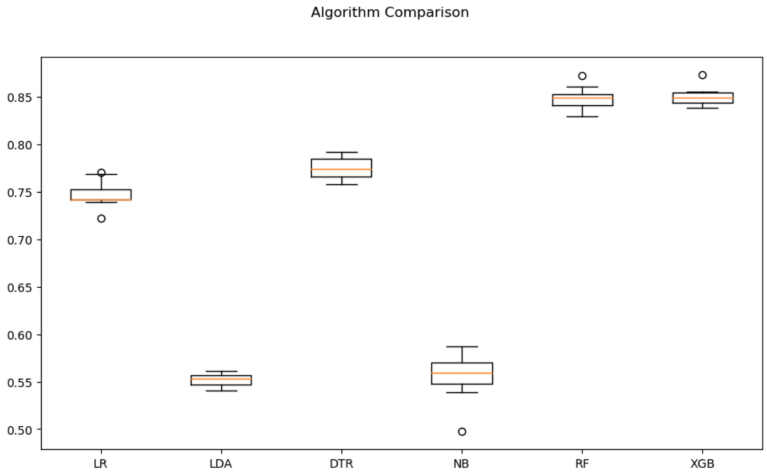


The number of not bankrupted companies is over 700k while the number of bankrupts is nearly 3k. The classification model trained and tested on the whole data would have low predictive quality due to the imbalance. Thus, the training sample containing not bankrupted companies has to be reduced extremely while all the bankrupted ones have to be saved. The new target value distribution of the training data is shown below:

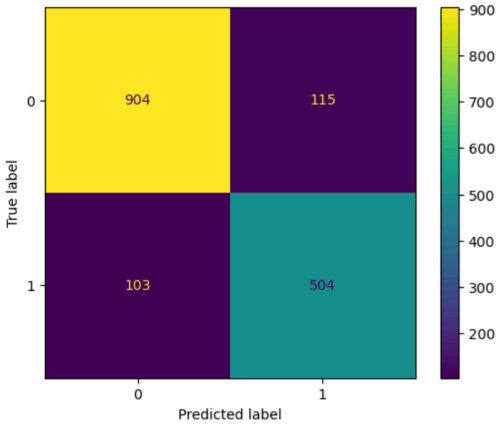


5k of not bankrupted companies are randomly chosen for the data balancing.

Different models are trained and tested using the cross-validation method to research the distributions of scores and find the most suitable one. Balanced accuracy score is chosen as a quality metric and the dataset is split 10 times. So, the graph containing box plots of scores of trained models is shown below:



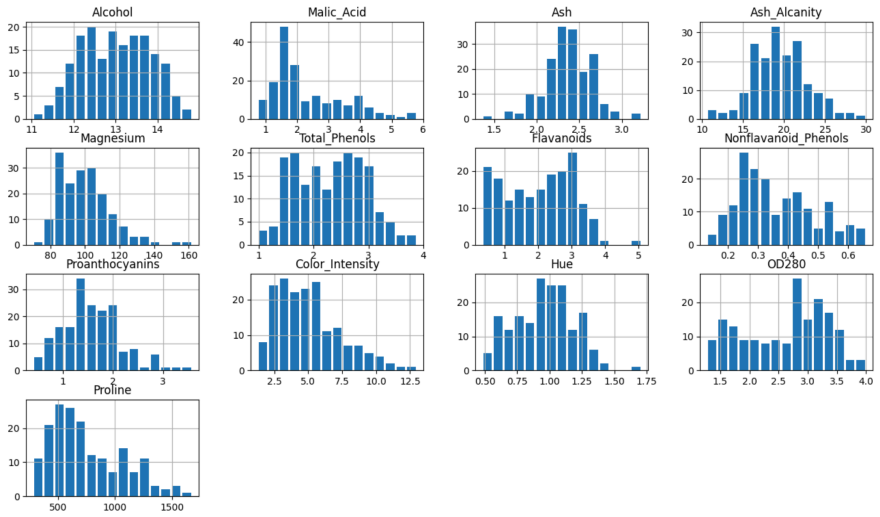
The ensemble models: XGB and Random Forest Classifier appear to be the most accurate ones while Logistic Regression and Decision Tree Classifier have moderately high mean balanced accuracies. Linear Discriminant Analysis and Naive Bayes show the worst results. The trained XGB shows balanced accuracy equal to 0,85 on a test set. Moreover, the confusion matrix below shows the distribution of the model predictions:



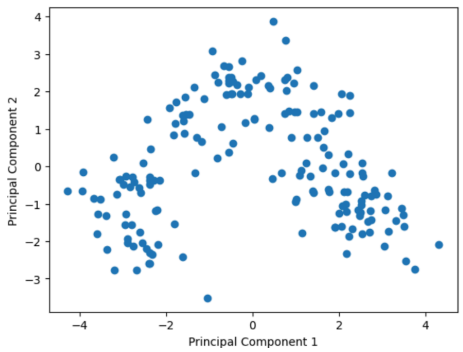
The quality of the model is high as it predicts the overwhelming majority of test samples correctly.

Unsupervised task: Wine dataset clustering

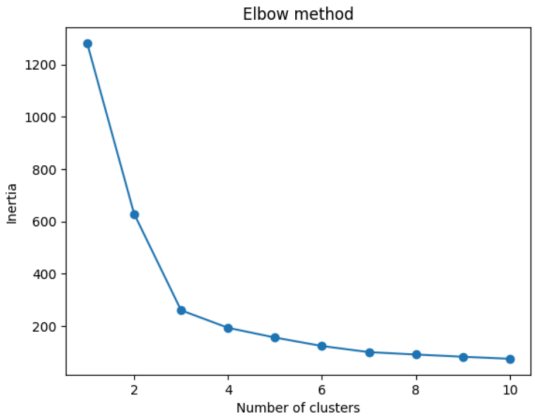
The wine dataset contains features of a chemical compound in every wine sample. However, It does not have class labels, and training a classification model and solving a classification task is impossible. Thus, the whole dataset has to be clusterised and the most suitable number of clusters has to be found. The dataset contains over 100 samples with only numerical features. The distributions of the features are shown below:

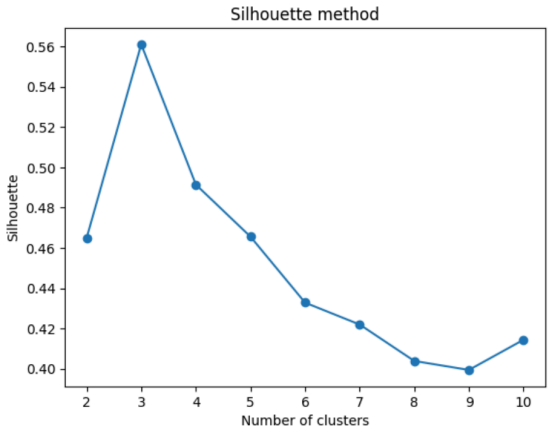


The feature transformation using a Standard Scaler will make the application of different models easier and more accurate. Thus, the new values are calculated using the next formula: z = (x - u) / s where «x» is an old value, «u» is the mean of a distribution, and «s» - standard deviation of a distribution. The reduction of 13 features to only 2 will help visualize and cluster the dataset. So, the method of Principal Component Analysis (PCA) is applied. It is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set (Jaadi, 2024). The new 2D dataset is visualised in a scatterplot graph below:

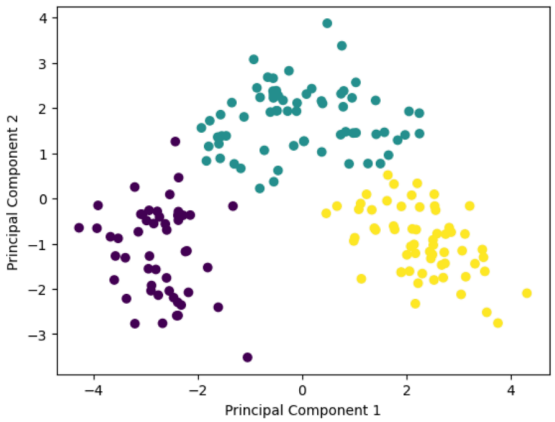


There are different methods of finding the most suitable number of clusters. The Elbow method is one of them. Every possible clustering of the dataset has a value of Inertia. Inertia is the sum of the squared distance of samples to their closest cluster centre (Herman-Saffar, 2021). We need the value of Inertia to be as small as possible, however, the number of clusters has to be optimal. So, the Elbow method helps to estimate a decrease in Inertia while adding new clusters. The Silhouette Analysis is another method of finding an optimal number of clusters. It computes silhouette coefficients of each point that measure how much a point is similar to its cluster compared to other clusters (Kumar, 2020). The value of the silhouette ranges between [1, -1], where a high value indicates that the object is well-matched to its cluster and poorly matched to neighbouring clusters. Thus, clustering of the decomposed dataset using the KMeans method with different numbers of clusters is made and both methods of finding the optimal number described previously are visualised below:





The Elbow method points number of clusters equal to 3 is optimal as adding new ones decreases inertia not significantly. Also, the model with 3 clusters has the highest value of the Silhouette coefficient. As it appears to be the most optimal number the scatterplot with 3 clusters is shown below:



References

1. Seoul Bike Sharing Demand. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5F62R.
2. Yeo, I. K., and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry.
3. HSE - Bankruptcy, Predict bankruptcy of companies using financial statements. (2024). Kaggle. https://www.kaggle.com/competitions/hse-m-psmsimmo-g-720353-1-bankruptcy-2018/overview
4. Aeberhard,Stefan and Forina,M.. (1991). Wine. UCI Machine Learning Repository. https://doi.org/10.24432/C5PC7J.
5. Jaadi, Z. (2024). A Step-by-Step Explanation of Principal Component Analysis (PCA). https://builtin.com/data-science/step-step-explanation-principal-component-analysis
6. Herman-Saffar, O. (2021). An Approach for Choosing Number of Clusters for K-Means. https://towardsdatascience.com/an-approach-for-choosing-number-of-clusters-for-k-means-c28e614ecb2c
7. Kumar, S. (2020). Silhouette Method — Better than Elbow Method to find Optimal Clusters. https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891