| FULL LEGAL NAME | **LOCATION (COUNTRY)** | **EMAIL ADDRESS** | **MARK X FOR ANY NON-CONTRIBUTING MEMBER** |
| --- | --- | --- | --- |
| Kumar Shantanu |  | Kumarshan25@gmail.com |  |
| Anton Dyadik |  | a.dyadik@gmail.com |  |
| Yu Li | Canada | coolsyd65@hotmail.com |  |

|  |  |
| --- | --- |
| **Statement of integrity:** By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an “X” above). | |
| **Team member 1** |  |
| **Team member 2** |  |
| **Team member 3** |  |

|  |
| --- |
| Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.  **Note:** You may be required to provide proof of your outreach to non-contributing members upon request. |
| N/A |

**Step 1 and 2:**

1. **Category 2 - K-Means clustering**
2. **Advantages**

* It is easy to understand as the goal is to segment data into different groups so that each group has similar data.
* It is easy to implement where large dataset can be quickly processed.
* It is flexible and adaptable as it can work with different types of datasets.
* It is robust and stable given the consistent results produced by the model.

1. **Basics**

* K-Means clustering attempts to group data with similar features into clusters.
* It is an unsupervised machine learning algorithm.

1. **Computation**

Refer to Jupyter notebook

1. **Disadvantages**

* Model results could be quite different as it is sensitive to the initial parameters.
* There is no clear rule regarding how to define parameter K. Different people might have different interpretation of K.
* It is sensitive to outlier/extreme value which may require preprocessing of the data to get normalized value.
* It can only work with numerical data because it needs to calculate the distances between data points, while categorical data cannot be measured with distance. Categorical value has to be converted into numerical value and this adds complexity to the model.

1. **Equations**

Source: Bustamam, A. et al. "Application of K-means Clustering Algorithm in Grouping the DNA Sequences of Hepatitis B Virus (HBV)." AIP Conference Proceedings, vol. 1862. no. 1. AIP Publishing LLC, 2017.

A diagram of a cluster

Description automatically generated

A white paper with black text and symbols

Description automatically generated

1. **Features**

* Unsupervised learning algorithm with no labeled data
* Train the model to find similarity of the data and then segment it into different categories/clusters
* Need to select the number of K at the beginning of the model by using either the Silhouette score or the elbow method.

1. **Guide**
   1. There are 5 steps to implement K-Means clustering:

* Step 1: Importing the libraries
* Step 2: Generating a dataset with 4 clusters from the scikit-learn library" make\_blobs" and plot it for visualization
* Step 3: Using the elbow method to find the optimal number of clusters
* Step 4: Training the K-Means model on the dataset
* Step 5: Visualizing the clusters
  1. Inputs:
* np.random.seed(0): setting the seed will ensure the same sequence of random numbers generated each time by the model.
* X, y = make\_blobs(centers=4, n\_samples=2500): generating 4 clusters from a dataset with 2500 points
* kmeans = KMeans(n\_clusters=i, init="k-means++", max\_iter=300, n\_init=10, random\_state=42): initializing the model’s parameters
  1. Outputs:

From the Elbow method, we can see that WCSS starts to decrease slowly around 4. Therefore, we will set K=4. K=4 is the output from elbow method.

1. **Hyperparameters**

* n\_clusters=i
* init
* max\_iter
* n\_init
* random\_state

1. **Illustration**

A diagram of colorful dots

Description automatically generated

A graph with a line

Description automatically generated

A diagram of different colored circles

Description automatically generated

1. **Journal:** "Application of Clustering Methods to Trading Strategies in the US Equity Market." *Imperial College London*, [www.imperial.ac.uk/media/imperial-college/faculty-of-natural-sciences/department-of-mathematics/math-finance/Lu\_Yilang\_01407813.pdf. Accessed 9 Dec. 2023](http://www.imperial.ac.uk/media/imperial-college/faculty-of-natural-sciences/department-of-mathematics/math-finance/Lu_Yilang_01407813.pdf.%20Accessed%209%20Dec.%202023).

The article has introduced three clustering methods and applied them in the US stock market to analyze 724 stocks from 10 sectors and 48 industries.

Although this article also covers other clustering algorithms, we only focus on K-Means clustering for the purpose of this assignment.

1. Firstly, the article provides the definition of unsupervised learning and a brief comparison of the differences between supervised learning and unsupervised learning with the latter focusing on identifying the hidden structure of input data.
2. Secondly, the author talked about the mathematics implication behind K-Means clustering and lists the following algorithms:
   * Determine initial centroids by selecting k variables from set S
   * Calculate distances between variables and centroids
   * Recalculate the centroid for each cluster and then repeat the prior step
   * Repeat the algorithm until we reach the optimal K
3. Thirdly, the article explains the result from applying the K-Means clustering where the algorithm only works well with sectors having good clustering property such as energy, financials, and utilities, while others without good clustering property are divided into several subsets and randomly distributed to many clusters.
4. Finally, the author concludes that K-means clustering does offer an incredible result which has improved the insider correlations of the clusters, but the weakness is that it can only find local optimal solutions.
5. **Keywords:** Tags to identify K-Means clustering: clustering, unsupervised learning, unlabeled, distance, Silhouette score, elbow method, centroid, hyperparameter

**Step 3: Technical Section**

1. **Parameters in each model**

**(1) K-Means clustering**

* n\_clusters
* init
* max\_iter
* n\_init
* random\_state

1. **Methods of hyperparameter tuning**

The following lists five ways of conducting hyperparameter tuning as well as their features.

(1) Manual checking: inefficient and time consuming

(2) K-fold cross validation: efficient and easy to use

(3) Grid search: suitable for exploring relatively few combinations

(4) Randomized search: suitable for searching large hyperparameters

(5) Ensemble method: combine several predictors from different trained models into a better predictor

1. **How to conduct hyperparameter tuning in each model**

**(1) K-Mean Clustering**

We will use grid search to find the optimal combination of hyperparameter values in K-Means clustering. The basic logic here is to search through all possible combinations of hyperparameter values using cross-validation to find the best performed hyperparameters. The following lists the two critical steps to conduct grid search:

1) We start by defining the parameter grid to search through.

In our model, we initially set the following parameter: kmeans = KMeans(n\_clusters=i, init="k-means++", max\_iter=300, n\_init=10, random\_state=42):

Now to perform grid search, we need to define our search parameters first:

param\_grid = {

'n\_clusters': [4, 5, 6, 7],

'init': ['k-means++', 'random'],

'max\_iter': [100, 200, 300],

'n\_init': [5, 10, 15],

'random\_state': [0, 42, 100]

}

2) We then use “GidSearchCV” from scikit-learn library to perform grid search with cross-validation.

grid\_search = GridSearchCV(kmeans, param\_grid, cv=5, scoring='silhouette')

Overall, grid search will help us identify the best combination of parameters.

**Step 4: Marketing Alpha**

1. **Applying K-Means clustering in retail sector**

**(1) Segmenting the data**

K-Means clustering is proved to be a fast and efficient way to find patterns from the large dataset for the purpose of grouping similar data together for better analysis. It can be used in retail sector to identify the customers with similarity so that a proper marketing strategy could be applied to each customer group to achieve better results.

For example, if we have the dataset of a shopping center with customer information of their name, annual income and annual spending, we can use K-Means clustering to find K clusters in the dataset where each cluster represents similar customers. We can see that K-Means cluster divides the customers into 4 groups which include high income and high spending, low income and low spending, high income and low spending, as well as low income and high spending.

**(2) Targeting the right customers-marketing alpha**

Continue with the above-mentioned example. Once we get the information about 4 different groups of customers, we can tailor different market strategies for each segment and achieve better results from our marketing campaign, i.e.: marketing alpha.

For example, marketing strategy for high income and low spending group should be tailored with creating an email list or newsletter to promote the brand, while strategy for low income and low spending should be focusing on promoting customer rewards as well as sales and discount activities.

**Step 5: More about machine learning**

**1. Machine learning in financial market**

A typical application of K-Means clustering is to partition stock price time series into different clusters to assist linear regression analysis for stock price prediction.

It can also be used to classify stocks into different sectors as illustrated by our referenced journal article related to K-Means clustering. The article identifies the anomaly where some stocks in different sectors tend to move together. This phenomenon poses the application of K-Means clustering so that we can better understand the relationship of the stock price in each sector.

The afore-mentioned two applications best explained the strengths of machine learning:

* Ability to process large and complex dataset such as financial data
* Fast and efficient to identify the patterns
* Moe accurate and reliable results
* Process automation could be achieved more easily

Overall, machine learning can help solve various financial problems to improve decision making process.

**Bibliography**

"Application of Clustering Methods to Trading Strategies in the US Equity Market." *Imperial College London*, [www.imperial.ac.uk/media/imperial-college/faculty-of-natural-sciences/department-of-mathematics/math-finance/Lu\_Yilang\_01407813.pdf. Accessed 9 Dec. 2023](http://www.imperial.ac.uk/media/imperial-college/faculty-of-natural-sciences/department-of-mathematics/math-finance/Lu_Yilang_01407813.pdf.%20Accessed%209%20Dec.%202023).

Bustamam, A. et al. "Application of K-means Clustering Algorithm in Grouping the DNA Sequences of Hepatitis B Virus (HBV)." AIP Conference Proceedings, vol. 1862. no. 1. AIP Publishing LLC, 2017.