**Dimension Reduction – PCA & SVD**

**Instructions:**

Please share your answers wherever applicable in line with the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: \_\_\_\_\_\_\_\_\_\_\_\_\_ Batch ID:** \_\_\_\_\_\_\_\_\_\_\_

**Topic: Dimension Reduction – PCA & SVD**

**Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) is submitted along with the documentation, explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered a correct submission.**

**2. Ensure that you submit your assignments correctly and in full. Resubmission is not allowed.**

**3. Post the submission you can evaluate your work by referring to the keys provided. (Will be available only post the submission).**

**Hints:**

**1. Business Problem**

* 1. **What is the business objective?**
  2. **What are the constraints?**
  3. **Define success criteria**

**2. Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Table

Description automatically generated

**3. Exploratory Data Analysis (EDA):**

**3.1. Univariate analysis.**

**3.2. Bivariate analysis.**

**4. Data Pre-processing**

**4.1 Data Cleaning, Feature Engineering, etc.**

**5.** **Multivariate Analysis**

**5.1 Build the model on the scaled data (try multiple options).**

**5.2 Perform the clustering and analyze the clusters.**

**5.3 Validate the clusters (try with the different numbers of clusters), label the clusters, and derive insights (compare the results from multiple approaches).**

**6. Use the clustered data and perform feature extraction using PCA and SVD. Compare the results.**

**7. Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**8. Deploy the best model using Python Flask on the local machine.**

**Problem Statements:**

The average retention rate in the insurance industry is 84%, with the top-performing agencies in the 93% - 95% range. Retaining customers is all about the long-term relationship you build. Offering a discount on the client’s current policy will ensure he/she buys a new product or renews the current policy. Studying clients' purchasing behaviour to determine which products they're most likely to buy is essential.

The insurance company wants to analyze their customer’s behaviour to strategies offers to increase customer loyalty.

**CRISP-ML(Q) process model describes six phases:**

1. Business and Data Understanding

2. Data Preparation

3. Model Building

4. Model Evaluation

5. Deployment

6. Monitoring and Maintenance

**Objective**: Maximize the Sales

**Constraints**: Minimize the Customer Retention

**Success Criteria:**

Business Success Criteria: Increase the Sales by 10% to 12% by targeting cross-selling opportunities on current customers.

ML Success Criteria: NA

Economic Success Criteria: The insurance company will see an increase in revenues by at least 8%

Data: Refer to the Autoinsurance.csv dataset.A picture containing chart

Description automatically generated

**Questions to Trigger Your thoughts:**

Q1. Which libraries are used in PCA to find the optimal number of PCA components?

Q2. Principal Component Analysis (PCA) is a \_\_\_\_\_\_\_\_\_\_ technique in Data Mining?

Q3. What is the importance of using PCA before the clustering?

Q4. Can we perform PCA on categorical features?

Q5. Why is it important to create pipelines?

Q6. Which libraries we can use to save or dump pipelines?

Q7. Why it is important to standardize the data in PCA?

## Q8. How can you obtain the principal components and the eigenvalues from Scikit-Learn PCA?

Q9. What is sklearn.pipeline extension used for?

Q10. Why do we use filterwarnings function? What library does it belong to and what are the uses of the library

Q11. What is the extension for the sklearn library to import TruncatedSVD?

Q12. How to read only the first 30 data rows?

Q13. What are the common functions used from the joblib library? Why do we use this library?

Q14. How to drop columns in location [5]

Q15. How to set the timeframe as an index?

Q16. How to check what imputation is better for replacing nan/infinity values?

Q17. What does figsize(x, y) define in plotting?

Q18. Can we define the type of plot inside a plot() function?

Q19. How is SVD different from PCA?

Q20. What are n\_components in SVD?

Q21. What does the fit function do? What does the transform function do?