Data Engineering

Week-2: Graded Project

Introduction:

Data processing and feature engineering are very important skills needed in data engineering. They have a lot of influence on the performance of machine learning models and even the quality of insights derived.

In this assignment, we will learn some important techniques and tools that will help you properly extract, prepare, and engineer features from your dataset.

Learning Objectives:

- Data analysis
- Dealing with missing values
- Checking for duplicates
- Dealing with categorical variables using One Hot Encoding and manual encoding
- Dropping irrelevant features
- PCA (Principal Component Analysis) for visualization and dimensionality reduction
- LDA for improving the model performance

Domain:

Marketing

Objective:

Visualizing high dimensional data using PCA and doing dimensionality reduction to check the explained variance using the PCA model. Training a Linear Discriminant Analysis(LDA) model to check if the product has been shipped or canceled.

Problem Statement:

XYZ.com is an e-commerce company based in Argentina. Due to the covid crisis and lockdown XYZ.com is facing lots of issues from the dealer and the shipment team. XYZ.com has lots of product data where various shipping and sales details of each product have been mentioned. XYZ.com wants to find out which of the products has been shipped and which of the products has been canceled to reduce customer escalation. As a data-scientist, we have to train a PCA model to visualize its higher-dimensional data and we have to train an LDA(Linear Discriminant Analysis) model to predict which of the products has been shipped and which of the products has been canceled.

Data Description:

The dataset can be found here.

Feature Details:

ORDERNUMBER: Order number of the product.

QUANTITYORDERED: Ordered quantity.

PRICEEACH: Price of each product.

ORDERLINENUMBER: Order line number of the product.

SALES: Sales of the product.

ORDERDATE:Order date of the product.

STATUS:Shipping status(i.e. Shipped or canceled or Resolved) (**TARGET**)

STATE: state where the product needs to be shipped

COUNTRY: Country where the product to be shipped.

DEALSIZE: Size of the product.

And so on…

The complete feature details can be found in the above mentioned link.

Tasks (A few steps are executed in the starter notebook):

Data Loading and Exploration.

1. Import necessary libraries.

2. Display a sample of five rows of the data frame.

3. Check the shape of the data (number of rows and columns). Check the general information about the dataframe using the .info() method.

4. Check the percentage of missing values in each column of the data frame. Drop the missing values if there are any.

5. Check if there are any duplicate rows.

6. Write a function that will impute missing values of the columns “STATE”, “POSTALCODE”,“TERRITORY” with its most occurring label.

7. Drop “ADDRESSLINE2”,”ORDERDATE”,”PHONE” column.

8. Convert the labels of the STATUS column to 0 and 1. For Shipped assign value 1 and for all other labels (i.e. ‘Cancelled’,’ Resolved’,’ On Hold’,’ In Process’, 'Disputed') assign 0. Note we will consider everything apart from Shipped as cancel (i.e. 0).

9. Assign ‘STATUS’ column into a label variable and drop it from the original dataframe.

10. Convert the original dataframe to the dummy coded data. (Hint:-use pd.get_dummies()

11. Use StandardScaler to scale the data.

- **PCA FOR VISUALIZATION**

1. Take the help of PCA to reduce the data to 2 dimensions. Use n_components=2.

2. Take the first and second principal components and plot a scatter plot with the labels.

3. Write the intuitions about the scatter plot.

- **PCA FOR DIMENSION REDUCTION.**

1. Fit the PCA model on the data and plot a graph between n_components and cumulative explained variance.

2. In how many components we are getting approximately 90% of explained variance.

- **LDA**

1. Split the dataset into two parts (i.e. 80% train and 20% test) using random_state=42.

2. Train a Linear Discriminant Analysis(LDA) model on the train data. Do fit_transform on the train data and only transform on the test data. Use n_components=1.

3. Train a RandomForest classifier model on the transformed train and test data. Print the accuracy score.

Submission

The final submission of this assessment should be made on Olympus. Make changes in the Jupyter Notebook and submit the same on Olympus in .ipynb format.