Big Data Project
Report
(Theory)

Surya Nikhil Mallampalli
AP20110010752

Vasireddy Pooja Vishali
AP20110010770

Gorantla Geetha Gayathri
AP20110010748

Konidela Pavan Rahul
AP20110010746

Tejaswini Pothuri
AP20110010728

INDEX

Dataset 1:SF Salaries
(San Francisco city employee salary data)

Source: https://www.kaggle.com/datasets/kaggle/sf-salaries

About Dataset:

One way to understand how a city government works is by looking at who it employs and how its employees are compensated. This data contains the names, job title, and compensation for San Francisco city employees on an annual basis from 2011 to 2014.

Used DataSet:
https://drive.google.com/file/d/1iTi5jSjxlwgbVSHz8JQYJatGzlxDF4dJ/view?usp=sharing

Approach Towards the Dataset:

Main Features in this dataset are **BasePay, OvertimePay, OtherPay, TotalPay, TotalPayBenefits,** and **Year** are numerical features.

So, we wanted to apply Preprocessing techniques on the TotalPay Feature of the dataset.They are as follows:

-Data Normalization: Min-Max normalization method is used for normalizing the feature and we can see that theTotalPay has min value 0 and max value 1.

-Outlier Removal: To improve the performance of the data set we remove the outliers(i.e) we have removed 1000 tuples from the 1lakh tuples.

-Log Transformation: Log Transformation is used to reduce the variability of the data .After performing the Log transformation on the TotalPay feature , we have reduced the standard deviation to 0.088905 from 50,0000.

Preprocess Code:

IPYNB File:

https://drive.google.com/file/d/12sPW3pWlRxYTz7EJpOg2QdCRcPr6H3qf/view?usp=sharing

.PY File

https://drive.google.com/file/d/1UGpIrQsnDyaPTejWUmqk7BHLVFNM8cZo/view?usp=sharing

Dataset 2:

(Credit Card Fraud Detection)

Source:https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

About Dataset:

The dataset contains transactions made by credit cards in September 2013 by European cardholders.
This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

Used DataSet:
https://drive.google.com/file/d/12bT-Y51CV8DIsdNxGi2SpQGJsolD329c/view?usp=sharing

Approach Towards the Dataset:

In this dataset, the data is labeled as 0's and 1's.
0-Not Fraud
1-Fraud

The Preprocessing Techniques applied here are:
-We checked if the dataset contained any NULL values and removed them if present.

-As the dataset is very huge we have only considered only 500 tuples from the dataset and it only has Not Fraud data
-As we have 30 Features in the dataset we plotted the heat map of the distribution of these features.
-Standardization of the features have been done without transforming the ‘Time’ and ‘Amount’ Features.
-Principal component analysis(PCA) is done to reduce the size of feature space by getting as much information as possible.It has been reduced to 2 features.
-Then we plotted a graph of Transaction class distribution against the Frequency.

Preprocess Code:

.IPYNB File:
https://drive.google.com/file/d/173166cT7Yt3bpORA9lnWNzgydggrLAru/view?usp=sharing
.PY File:
https://drive.google.com/file/d/1ibYwsGNh-cdPhLONn43X6b8WEKBfQT4Z/view?usp=sharing

Dataset 3:Hotel Dataset

Source:https://www.kaggle.com/competitions/expedia-hotel-recommendations/data

About Dataset:

This dataset contains hotel bookings on different sites.

Used Dataset: https://drive.google.com/drive/folders/1opGboKcgtnNGx-67mkaKBkQ_QRZ_92pU?usp=sharing

Approach Towards the Dataset:

To clean and pre-process the data and perform exploratory analysis to get some interesting insights into the process of choosing a hotel.

1. Remove the users who did not booked the hotel
2. Identify the searches by each user belonging to a specific type of destination
3. orig_destination_distance contains Nan values
4. The check-in and check-out dates to find the duration of the stay for each of the entries in the training set.

- In the data cleaning process we added Additional features like date columns(day/month/year) and Attributes like stay duration,no_of_days between bookings,Check-in day/month/year.

-Then we convert date objects to relevant attributes and fill in the missing values by training the dataset.And it fills the avg values in place of nan with mean values.

-Then we remove the unnecessary objects from the dataset.

Preprocess Code:

.IPYNB File:
https://drive.google.com/file/d/1cb0-Cblf_6lvDhwMtkgTHzxZUuM0Q3Qe/view?usp=sharing

.PY File:
https://drive.google.com/file/d/1611rz5vakYI_EkFPfNcwWhy5-L4Wx7OW/view?usp=sharing