Lab Assignment 1

CIS 660 Data Mining

Adventure Works Cycles

**Data Processing Language : Python 3.5 IDE**

Part 1: Feature Selection, Cleaning, and Preprocessing to Construct an Input from Data Source

-Purchase a bike is not effected by name of customer. So, droping that columns.

-Also many states and cities to work on. So, droping all address related data and only using country for getting results

CustomerKey Discrete

GeographyKey Discrete

CustomerAlternateKey Nominal

Gender Nominal

MaritalStatus Nominal

EnglishEducation Nominal

SpanishEducation Nominal

FrenchEducation Nominal

EnglishOccupation Nominal

SpanishOccupation Nominal

FrenchOccupation Nominal

HouseOwnerFlag Discrete

DateFirstPurchase Nominal

CommuteDistance Nominal

Region Nominal

Age Discrete

BikeBuyer Discrete

NumberCarsOwned Discrete

NumberChildrenAtHome Discrete

TotalChildren Discrete

YearlyIncome Discrete

Part 2: Data Preprocessing and Transformation

One Hot Encoding for nominal data

1)All Education

2) All occupation

Part 3: Calculating Proximity of Two Binary Object Vectors With Simple Matching , Jaccard Similarity, Cosine Similarity

EnglishEducation Nominal

EnglishOccupation Nominal

HouseOwnerFlag Discrete

DateFirstPurchase Nominal

CommuteDistance Nominal

Region Nominal

Age Discrete

BikeBuyer Discrete

NumberCarsOwned Discrete

NumberChildrenAtHome Discrete

TotalChildren Discrete

YearlyIncome Discrete

transforming the values in Normalizer form

applied

Jaccard Similarity . Cosine Similarity and pearsonr

for occupation and YearlyIncome ------------------(1)

and also for Education and YearlyIncome --------(2)

which gave the values

Cosine Similarity

(1):0.42748354460006355

(2):0.5826049619920608

Jaccard Similarity

(1):1.0

(2):1.0

pearsonr

(1):0.4845560071281075

(2):0.12432250951401452