TEAM Galaxy

(Prashant Raghuwanshi & Jay Pfister)

DSC630-T301 Predictive Analytics (2223-1)

Professor Fadi Alsaleem

Milestone 4, Project Proposal

01/29/2022

Each phase of the process:

Template Source : https://www.sv-europe.com/crisp-dm-methodology/

Abstract:

-This term-end project aims to evaluate the Students should be able to identify a business problem to address through predictive analytics. The goal is to select appropriate models and model specifications and apply the respective methods to enhance data-driven decision-making related to the business problem.

Students will identify the potential use of predictive analytics, formulate the problem, identify the right sources of data, analyze data and prescribe actions to improve not only the process of decision making but also the outcome of decisions.

1. Stage One - Determine Business Objectives and Assess the Situation

1.1 Assess the Current Situation

We are targeting a sub set of survey respondents who we have identified as gainfully and regularly employed and potentially in need of financial services such as savings accounts or CDs.

Business Application - The goal of this project is to assist in the marketing of financial services to customers based on self identified demographic data from the Bureau of Labor Statistics, allowing for the creation of pools of demographically sorted candidates for client services. Based on our findings, we can craft tailored survey data to better target ideal customers for our clients. This goal is met by creating a predictive model to determine if a given group of respondents is likely to be a good fit for our clients' financial products and as a result an ideal target for marketing campaigns attempting to solicit customers for those products.

1.1.1. Inventory of resources

List the resources available to the project including:

- Personnel: Prashant, Jay
- Data: ATUS 2019 Respondent Data
- Computing resources: Personal computers
- Software: Jupyter Notebook(Python), R Studio, IBM SPSS

1.1.2. Requirements, assumptions and constraints -

- Major Requirements: A valid a predictive variable to determine viability of a candidate based on survey response in order to validate the output of our predictive model that would then enable us to target specific demographics based on respondent inputs.
- A pool of survey respondents providing demographic data about their employment status as well as details about their willingness to work and relevant experience (previous employment related responses).
- Major Constraints: A possible constraint is a signifcant number of declined responses from those surveyed. Without enough input, we won't be able to make a meaningful prediction based on the data, or lack thereof.
- The model also relies on a candidate's real time behavior, rather than a prepared list of responses. Because of this, there is a need for respondents to respond to the survey candidly and truthfully.

1.1.3.Risks and contingencies

- Risks include potential PII issues with respondents, which can largely be mitigated by anonymized respondent IDs. As long as a particular individual is never personally idenitified the risk of PII information leaking is mitigated.
- Potential for respondents to misrepresent their individual situations in their survey replies. Survey responses are to some extent unverifiable. For example, if a respondent indicates that the reason they remain unemployed is a long term medical issue, we can't verify that claim and that could potentially skew results.
- Lack of suitable candidates for a given application based on survey responses. This can be mitigated by the model indicating viability within the candidate pool.

1.1.4.Terminology

CRISP-DM The Cross-Industry Standard Process for Data-Mining – CRISP-DM is a model of a data mining process used to solve problems by experts. The model identifies the different stages in implementing a data mining project, as described bellow The model proposes the following steps: • Business Understanding – to understand the rules and business objectives of the company. • Understanding Data – to collect and describe data. • Data Preparation – to prepare data for import into the software. • Modeling – to select the modeling technique to be used. • Evaluation – Evaluate the process to see if the technique solves modeling and creating rules. • Deployment – to deploy the system and train its users

1.1.5.Costs and benefits

- Costs:
- Primary cost associated with this project is the time of the people working on it.
- Computing resources for modeling

- Benefits:

- Social benefit of assisting the process of filling employment gaps, and potentially enabling otherwise unemployable people to find employment
- Financial benefit to the company from the ability to sell this product to the target consumer.

1.2 What are the desired outputs of the project?

Business success criteria

- Provide this information to financial institutions for targeting for financial products.

Data mining success criteria

- Finding predictive demographic data from the respondent pool in order to determine what variables are best used for candidate selection.

Produce project plan

- Adhere to CRISP-DM methodology. Data prep Understanding > Data Prep > Modelling > Evaluation > Deployment

progress.png

1.3 What Questions Are We Trying To Answer?

Can we identify survey respondent segments that are candidates for potential packaging as demographic data to sell to our customers?

What traits make an individual most suitable for the financial services for our customers?

Which of the survey respondents are good fits for the financial products of our customers?

2. Stage Two - Data Understanding

The second stage of the CRISP-DM process requires you to acquire the data listed in the project resources. This initial collection includes data loading, if this is necessary for data understanding. For example, if you use a specific tool for data understanding, it makes perfect sense to load your data into this tool. If you acquire multiple data sources then you need to consider how and when you're going to integrate these.

- Our Data set is a survey dataset and it contains the question related to two sections,
- The first section is income/employment
- The second section is related to time use

2.1 Initial Data Report

Initial data collection report - List the data sources acquired together with their locations, the methods used to acquire them and any problems encountered. Record problems you encountered and any resolutions achieved. This will help both with future replication of this project and with the execution of similar future projects.

In [1]:
```python
# Import Libraries Required
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import numpy as np
import seaborn as sns
pd.set_option("display.max_columns", None)
from scipy.cluster.hierarchy import dendrogram
from sklearn.cluster import AgglomerativeClustering
from sklearn.preprocessing import StandardScaler, normalize
from sklearn.metrics import silhouette_score
import scipy.cluster.hierarchy as shc
from sklearn.decomposition import PCA
# supress warnings
import warnings
import seaborn as sns
from matplotlib.pyplot import xticks
from sklearn import preprocessing
from kmodes.kmodes import KModes
```

In [2]:
```python
#Data source:
#Source Query location:
path =  '2019 BLS.csv'
# reads the data from the file - denotes as CSV, it has no header, sets column headers
df =  pd.read_csv(path, sep=',')
df0 = df.copy()
```

2.2 Describe Data

Data description report - ATUS Respondent File This file contains case-specific variables collected in ATUS (that is, variables for which there is one value for each respondent). These include, for example, labor force and earnings information, total time providing secondary childcare, and ATUS statistical weights.

There is one record for each ATUS respondent.

Below is a simplified example. The variable TUCASEID identifies each household, and the variable TULINENO identifies each individual within the household. The example contains responses from five individuals; note that the respondent always has TULINENO = 1. In the example, each respondent has a corresponding statistical weight for use in generating estimates representative of the U.S. civilian, noninstitutionalized population (TUFINLWGT), and values for school enrollment (TESCHENR), labor force status (TELFS), and total number of minutes spent alone on the diary day (TRTALONE). The actual ATUS Respondent file contains more variables and records.

TUCASEID	TULINENO	TUFINLWGT	TESCHENR	TELFS	TRTALONE
20060101020210	1	22261358.19	1	1	40
20060101020211	1	5019645.31	1	1	350
20060101020212	1	2926068.74	1	5	0
20060101020213	1	25780574.07	2	5	556
20060101020214	1	3414645.94	1	4	100

Income & Employment:

TEERNHRY "Edited: hourly-non-hourly status" / TEERNPER "Edited: for your main job, what is the easiest way for you to report your total earnings before taxes or other deductions: hourl" /
TEERNUOT "Edited: do you usually receive overtime pay, tips, or commissions at your main job?" / TEHRUSL1 "Edited: how many hours per week do you usually work at your main job?" / TEHRUSLT "Edited: total hours usually worked per week [sum of TEHRUSL1 and TEHRUSL2]" / TEIO1COW "Edited: individual class of worker code [main job]" / TEIO1ICD "Edited: industry code [main job]" / TEIO1OCD "Edited: occupation code [main job]" / TELFS "Edited: labor force status" /
TEMJOT "Edited: in the last seven days did you have more than one job?" / TESPEMPNOT "Edited: employment status of spouse or unmarried partner" / TRDPFTPT "Full time or part time employment status of respondent" / TRDTIND1 "Detailed industry recode [main job]" / TRDTOCC1 "Detailed occupation recode [main job]" / TRIMIND1 "Intermediate industry recode [main job]" / TRMJIND1 "Major industry recode [main job]" / TRMJOCC1 "Major occupation recode [main job]" / TRMJOCGR "Major occupation category [main job]" / TTWK "Weekly earnings topcode flag" / TUIO1MFG "Is this business or organization mainly manufacturing, retail trade, wholesale trade, or something else? [main job]"/ TUIODP1 "Last time we spoke to someone in this household, you were reported to work for [employer's name]. TUSPWK "In the last seven days, did your spouse or unmarried partner do any work for pay or profit?" / TUBUS "Does anyone in the household own a business or a farm?" /

In [3]:
```python
df.columns
```

Out[3]:
```
Index(['TUCASEID', 'TULINENO', 'TUYEAR', 'TUMONTH', 'TEABSRSN', 'TEERN',
       'TEERNH1O', 'TEERNH2', 'TEERNHRO', 'TEERNHRY',
       ...
       'TXSPEMPNOT', 'TXSPUHRS', 'TXTCC', 'TXTCCTOT', 'TXTCOC', 'TXTHH',
       'TXTNOHH', 'TXTO', 'TXTOHH', 'TXTONHH'],
      dtype='object', length=174)
```

In [4]:
```python
df.shape
```

Out[4]:
```
(9435, 174)
```

In [5]:
```python
df.dtypes
```

Out[5]:
```
TUCASEID    int64
TULINENO    int64
TUYEAR      int64
TUMONTH     int64
TEABSRSN    int64
            ...
TXTHH       int64
TXTNOHH     int64
TXTO        int64
TXTOHH      int64
TXTONHH     int64
Length: 174, dtype: object
```

In [6]:
```python
df.describe()
```

Out[6]:

	TUCASEID	TULINENO	TUYEAR	TUMONTH	TEABSRSN	TEERN	TEERNH1O	TEERNH2	TEERNHRO	TEERNHRY	TEER
count	9.435000e+03	9435.0	9435.0	9435.000000	9435.00000	9435.000000	9435.000000	9435.000000	9435.000000	9435.000000	9435.0
mean	2.019065e+13	1.0	2019.0	6.404981	-0.78124	948.019714	356.827981	221.663911	5.537573	0.337785	1.5
std	3.496214e+08	0.0	0.0	3.478020	1.42014	7882.777017	883.983735	811.009246	14.481475	1.286188	2.8
min	2.019010e+13	1.0	2019.0	1.000000	-1.00000	-1.000000	-1.000000	-1.000000	-2.000000	-1.000000	-1.0
25%	2.019030e+13	1.0	2019.0	3.000000	-1.00000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.0
50%	2.019061e+13	1.0	2019.0	6.000000	-1.00000	-1.000000	-1.000000	-1.000000	-1.000000	1.000000	1.0
75%	2.019091e+13	1.0	2019.0	9.000000	-1.00000	-1.000000	-1.000000	-1.000000	-1.000000	2.000000	5.0
max	2.019121e+13	1.0	2019.0	12.000000	14.00000	225000.000000	9999.000000	9999.000000	99.000000	2.000000	7.0

In [7]:
```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9435 entries, 0 to 9434
Columns: 174 entries, TUCASEID to TXTONHH
dtypes: float64(1), int64(171), object(2)
memory usage: 12.5+ MB
```

In [8]:
```python
df.head(5)
```

Out[8]:

	TUCASEID	TULINENO	TUYEAR	TUMONTH	TEABSRSN	TEERN	TEERNH1O	TEERNH2	TEERNHRO	TEERNHRY	TEERNPER	TEERNRT
0	20190101190022	1	2019	1	-1	-1	-1	-1	-1	-1	-1	-1
1	20190101190031	1	2019	1	-1	-1	1100	-1	15	1	1	-1
2	20190101190057	1	2019	1	-1	-1	-1	1500	-1	1	2	1
3	20190101190107	1	2019	1	-1	-1	-1	-1	-1	-1	-1	-1
4	20190101190555	1	2019	1	-1	-1	-1	-1	-1	-1	-1	-1

In [9]:
```python
#splitting the data sets on required inclome & employment subject area .
df = df[["TEERNHRY", "TEERNPER", "TEERNUOT", "TEHRUSLT", "TEIO1COW", "TEIO1ICD", "TELFS", "TEMJOT", "TESPEMPNOT
```

2.3 Verify Data Quality

Examine the quality of the data, addressing questions such as:

- Is the data complete (does it cover all the cases required)?
- Is it correct, or does it contain errors and, if there are errors, how common are they?
- Are there missing values in the data? If so, how are they represented, where do they occur, and how common are they?

2.3.1. Missing Data

In addition to incorrect datatypes, another common problem when dealing with real-world data is missing values. These can arise for many reasons and have to be either filled in or removed before we train a machine learning model. First, let's get a sense of how many missing values are in each column

While we always want to be careful about removing information, if a column has a high percentage of missing values, then it probably will not be useful to our model. The threshold for removing columns should depend on the problem

In [10]:
```python
# checking null value
df.isnull().sum()
```

Out[10]:
```
TEERNHRY      0
TEERNPER      0
TEERNUOT      0
TEHRUSLT      0
TEIO1COW      0
TEIO1ICD      0
TELFS         0
TEMJOT        0
TESPEMPNOT    0
TRDPFTPT      0
TRDTIND1      0
TRDTOCC1      0
TRIMIND1      0
TRMJIND1      0
TRMJOCC1      0
TRMJOCGR      0
TUFWK         0
TUIO1MFG      0
TUIODP1       0
TUSPWK        0
TUBUS         0
dtype: int64
```

In [11]:
```python
# finding the columns which contains more missing values (greater than 50% of total records count)
# if data valuye contains -1. -2 , -3 entry then it means the user havent responded the survey question
a = [-1, -2, -3]
pct_null = df.isin(a).sum() / len(df)
missing_features = pct_null[pct_null > 0.50].index
missing_features
```

Out[11]:
```
Index([], dtype='object')
```

In [12]:
```python
# dropping the missing value columns
df.drop(missing_features, axis=1, inplace=True)
```

In [13]:
```python
# filling out missing values
df = df.fillna(method="ffill")
df = df.fillna(method="bfill")
```

In [14]:
```python
df.shape
```

Out[14]:
```
(9435, 21)
```

2.3.2 Removing Duplicate columns

In [15]:
```python
# This function take a dataframe
# as a parameter and returning list
# of column names whose contents
# are duplicates.
def getDuplicateColumns(df):

    # Create an empty set
    duplicateColumnNames = set()

    # Iterate through all the columns
    # of dataframe
    for x in range(df.shape[1]):

        # Take column at xth index.
        col = df.iloc[:, x]

        # Iterate through all the columns in
        # DataFrame from (x + 1)th index to
        # last index
        for y in range(x + 1, df.shape[1]):

            # Take column at yth index.
            otherCol = df.iloc[:, y]

            # Check if two columns at x & y
            # index are equal or not,
            # if equal then adding
            # to the set
            if col.equals(otherCol):
                duplicateColumnNames.add(df.columns.values[y])

    # Return list of unique column names
    # whose contents are duplicates.
    return list(duplicateColumnNames)
```

In [16]:
```python
drop_columns2= getDuplicateColumns(df)
# list duplicate data columns
drop_columns2
```

Out[16]:
```
[]
```

In [17]:
```python
# drop columns from df whose of values are duplicate
df.drop(drop_columns2, axis=1, inplace=True)
```

In [18]:
```python
def rep_data(df1):
    for col in df1.columns:
        df1[col] = df1[col].map( lambda x : df1[col].median() if x == -1 else x)
        df1[col] = df1[col].map( lambda x : df1[col].median() if x == -2 else x)
        df1[col] = df1[col].map( lambda x : df1[col].median() if x == -3 else x)
        df1[col] = df1[col].map( lambda x : df1[col].median() if x == -4 else x)
```

In [19]:
```python
# replacing missing value with median
rep_data(df)
```

3. Stage Three - Data Preperation

This is the stage of the project where you decide on the data that you're going to use for analysis. The criteria you might use to make this decision include the relevance of the data to your data mining goals, the quality of the data, and also technical constraints such as limits on data volume or data types. Note that data selection covers selection of attributes (columns) as well as selection of records (rows) in a table.

3.1 Label Encoding

In [20]:
```python
df.dtypes
```

Out[20]:
```
TEERNHRY      float64
TEERNPER      float64
TEERNUOT      float64
TEHRUSLT      float64
TEIO1COW      float64
TEIO1ICD      float64
TELFS           int64
TEMJOT        float64
TESPEMPNOT    float64
TRDPFTPT      float64
TRDTIND1      float64
TRDTOCC1      float64
TRIMIND1      float64
TRMJIND1      float64
TRMJOCC1      float64
TRMJOCGR      float64
TUFWK         float64
TUIO1MFG      float64
TUIODP1       float64
TUSPWK        float64
TUBUS         float64
dtype: object
```

In [21]:
```python
# Creating copy data set
df_employment = df.copy()
```

In [22]:
```python
df_employment1 = df_employment.copy()
```

3.2 One Hot Encoding

- Here we used One Hot Encoding to help with our largely categorical data set. The purpose of One Hot Encoding is to map various categorical variables like the ones in our data to integers that can be better used and understood by machine learning algorithms. This is particularly useful when working with data that may not all be related. It also helps when working with integers because of the way machine learning treats number order (Higher is better), as well as allowing for better scalability.

In [23]:
```python
# get wholeset of dummy variable for all columns
def one_hot_top_x(df2, variable, top_x_labels):
    for label in top_x_labels:
        df2[variable+'_'+str(label)] = np.where(df2[variable] == label, 1, 0)
```

In [24]:
```python
# hot encoding for top 5 variables from each columns
top_5 = [x for x in df_employment1.TEERNPER.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TEERNPER', top_5)
top_5 = [x for x in df_employment1.TEHRUSLT.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TEHRUSLT', top_5)
top_5 = [x for x in df_employment1.TEIO1COW.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TEIO1COW', top_5)
top_5 = [x for x in df_employment1.TEIO1ICD.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TEIO1ICD', top_5)
top_5 = [x for x in df_employment1.TELFS.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TELFS', top_5)
top_5 = [x for x in df_employment1.TRDTIND1.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TRDTIND1', top_5)
top_5 = [x for x in df_employment1.TRDTOCC1.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TRDTOCC1', top_5)
top_5 = [x for x in df_employment1.TRIMIND1.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TRIMIND1', top_5)
top_5 = [x for x in df_employment1.TRMJIND1.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TRMJIND1', top_5)
top_5 = [x for x in df_employment1.TRMJOCC1.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TRMJOCC1', top_5)
top_5 = [x for x in df_employment1.TRDTOCC1.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TRDTOCC1', top_5)
top_5 = [x for x in df_employment1.TRMJOCGR.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TRMJOCGR', top_5)
top_5 = [x for x in df_employment1.TEERNHRY.value_counts().sort_values(ascending=False).head(2).index]
one_hot_top_x(df_employment1, 'TEERNHRY', top_5)
top_5 = [x for x in df_employment1.TRMJOCGR.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TRMJOCGR', top_5)
top_5 = [x for x in df_employment1.TUFWK.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TUFWK', top_5)
top_5 = [x for x in df_employment1.TUIO1MFG.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TUIO1MFG', top_5)
top_5 = [x for x in df_employment1.TUIODP1.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TUIODP1', top_5)
top_5 = [x for x in df_employment1.TUSPWK.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TUSPWK', top_5)
top_5 = [x for x in df_employment1.TUBUS.value_counts().sort_values(ascending=False).head(5).index]
one_hot_top_x(df_employment1, 'TUBUS', top_5)
```

In [25]:
```python
# drop the main columns
dcol = ['TEERNHRY', 'TEERNPER', 'TEERNUOT', 'TEHRUSLT', 'TEIO1COW', 'TEIO1ICD', 'TELFS', 'TEMJOT', 'TESPEMPNOT'
df_employment2 = df_employment1.drop(columns = dcol,  axis = 1)
df_employment2.shape
```

Out[25]:
```
(9435, 75)
```

In [26]:
```python
# validating hot encoding
#for cols in df_employment2.columns:
#     print(cols, len(df_employment2[cols].unique()), df_employment2[cols].unique())
```

3.2.2 Drop Unnecessary Columns

- Sometimes we may not need certain columns. We can drop extra columns in order to keep only relevent data and avoid potentially spurious associations.

In [27]:
```python
# find out columns with bias data values
dtl2 = []
def bias_data1(df1):
    for coll in df1.columns:
        dtl = df1[coll].value_counts()
        dt3 = pd.DataFrame(dtl)
        dtl2.append(dt3/9435)
        dtl = ''
```

In [28]:
```python
bias_data1(df_employment2)
# after analysing the biased data results, we found below columns contains the biased categorical data vaues (
#['TEERNPER_2.0', 'TEERNPER_3.0', 'TEERNPER_5.0', 'TEHRUSLT_50.0', 'TEHRUSLT_60.0', 'TEIO1COW_5.0', 'TEIO1COW_
#       'TEIO1COW_7.0', 'TEIO1COW_2.0', 'TEIO1ICD_7860.0', 'TEIO1ICD_770', 'TEIO1ICD_8190.0', 'TEIO1ICD_8680.0
#       'TUSPWK_4.0', 'TUSPWK_5.0', 'TUIO1MFG_3.0', 'TUFWK_5.0', 'TUFWK_3.0', 'TELFS_3', 'TELFS_4']
```

In [29]:
```python
dcols =  ['TEERNPER_2.0', 'TEERNPER_3.0', 'TEERNPER_5.0', 'TEHRUSLT_50.0', 'TEHRUSLT_60.0', 'TEIO1COW_5.0', 'TE
       'TEIO1COW_7.0', 'TEIO1COW_2.0', 'TEIO1ICD_7860.0', 'TEIO1ICD_8190.0', 'TEIO1ICD_8680.0', 'TELFS_2', 'TEI
       'TUSPWK_4.0', 'TUSPWK_5.0', 'TUIO1MFG_3.0', 'TUFWK_5.0', 'TUFWK_3.0', 'TELFS_3', 'TELFS_4']
df_employment3 = df_employment2.drop(columns = dcols,  axis = 1)
df_employment3.shape
```

Out[29]:
```
(9435, 54)
```

In [30]:
```python
#df_employment3.head()
```

3.2 Dealing With Zeros

- Replacing all the zeros from columns.

- Zero values were previously addressed using mean value imputation. Some of our data contained 0's or other indicated non values. We chose to use Mean Value Imputation to account for these zeroes. Mean Value Imputation does exactly what it sounds like, it replaces the non-values in our data, for example 0, with the mean value of all the values in the column the data is found in.

4. Stage Four - Exploratory Data Analysis

4.1. Outliers

At this point, we may also want to remove outliers. These can be due to typos in data entry, mistakes in units, or they could be legitimate but extreme values. For this project, we will remove anomalies based on the definition of extreme outliers:

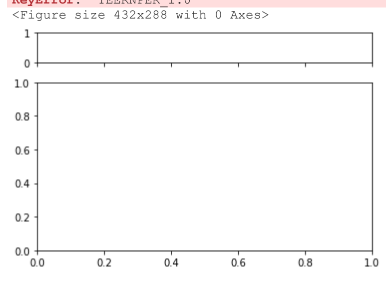
https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

- Below the first quartile − 3 ∗ interquartile range
- Above the third quartile + 3 ∗ interquartile range

In [31]:
```python
# showing boxplot & histogram of each features before removing outliers
for column in df_employment3:
    plt.figure()
    # creating a figure composed of two matplotlib.Axes objects (ax_box and ax_hist)
    f, (ax_box, ax_hist) = plt.subplots(2, sharex=True, gridspec_kw={"height_ratios": (.15, .85)})

   # assigning a graph to each ax
    sns.boxplot(df[column], ax=ax_box)
    sns.histplot(data=df, x=column, ax=ax_hist)

    # Remove x axis name for the boxplot
    ax_box.set(xlabel='')
    # comment out break statment to generate plots for al features
    break
plt.show()
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
~\AppData\Roaming\Python\Python38\site-packages\pandas\core\indexes\base.py in get_loc(self, key, method, toler
ance)
   3360             try:
-> 3361                 return self._engine.get_loc(casted_key)
   3362             except KeyError as err:

~\AppData\Roaming\Python\Python38\site-packages\pandas\_libs\index.pyx in pandas._libs.index.IndexEngine.get_lo
c()

~\AppData\Roaming\Python\Python38\site-packages\pandas\_libs\index.pyx in pandas._libs.index.IndexEngine.get_lo
c()

pandas\_libs\hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.get_item()

pandas\_libs\hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.get_item()

KeyError: 'TEERNPER_1.0'

The above exception was the direct cause of the following exception:

KeyError                                  Traceback (most recent call last)
<ipython-input-31-5208298aa5e7> in <module>
      6
      7     # assigning a graph to each ax
----> 8     sns.boxplot(df[column], ax=ax_box)
      9     sns.histplot(data=df, x=column, ax=ax_hist)
     10

~\AppData\Roaming\Python\Python38\site-packages\pandas\core\frame.py in __getitem__(self, key)
   3456             if self.columns.nlevels > 1:
   3457                 return self._getitem_multilevel(key)
-> 3458             indexer = self.columns.get_loc(key)
   3459             if is_integer(indexer):
   3460                 indexer = [indexer]

~\AppData\Roaming\Python\Python38\site-packages\pandas\core\indexes\base.py in get_loc(self, key, method, toler
ance)
   3361                 return self._engine.get_loc(casted_key)
   3362             except KeyError as err:
-> 3363                 raise KeyError(key) from err
   3364
   3365         if is_scalar(key) and isna(key) and not self.hasnans:

KeyError: 'TEERNPER_1.0'
```

```
<Figure size 432x288 with 0 Axes>
```



- After doing hot encoding & removing columns with bias records,
- Now all of our columns contain 0 or 1, so we dont required outlayer deduction and historgram plot

4.2 Initial Data Exploration

During this stage you'll address data mining questions using querying, data visualization and reporting techniques. These may include:

- **Distribution** of key attributes (for example, the target attribute of a prediction task)
- **Relationships** between pairs or small numbers of attributes
- Results of **simple aggregations**
- **Properties** of significant sub-populations
- **Simple** statistical analyses

These analyses may directly address your data mining goals. They may also contribute to or refine the data description and quality reports, and feed into the transformation and other data preparation steps needed for further analysis.

- **Data exploration report** - Describe results of your data exploration, including first findings or initial hypothesis and their impact on the remainder of the project. If appropriate you could include graphs and plots here to indicate data characteristics that suggest further examination of interesting data subsets.

4.2.1 Distributions

In [ ]:
```python
df_employment2.describe()
```

4.2.2 Correlations

Can we derive any correlation from this data-set? Pairplot chart gives us correlations, distributions and regression path Correlogram are awesome for exploratory analysis. It allows to quickly observe the relationship between every variable of your matrix. It is easy to do it with seaborn: just call the pairplot function

Pairplot Documentation can be found here: https://seaborn.pydata.org/generated/seaborn.pairplot.html

In [ ]:
```python
#corr = df_employment.apply(lambda x : pd.factorize(x)[0]).corr(method='pearson', min_periods=1)
#corr.style.background_gradient(cmap='coolwarm')

import numpy as np
import pandas as pd
import scipy.stats as ss
import seaborn as sns

def cramers_v(confusion_matrix):
    """ calculate Cramers V statistic for categorial-categorial association.
        uses correction from Bergsma and Wicher,
        Journal of the Korean Statistical Society 42 (2013): 323-328
    """
    chi2 = ss.chi2_contingency(confusion_matrix)[0]
    n = confusion_matrix.sum()
    phi2 = chi2 / n
    r, k = confusion_matrix.shape
    phi2corr = max(0, phi2 - ((k-1)*(r-1))/(n-1))
    rcorr = r - ((r-1)**2)/(n-1)
    kcorr = k - ((k-1)**2)/(n-1)
    return np.sqrt(phi2corr / min((kcorr-1), (rcorr-1)))
```

In [ ]:
```python
factors_paired = [(i,j) for i in df_employment2.columns.values for j in df_employment2.columns.values]

chi2, p_values =[], []

for f in factors_paired:
    if f[0] != f[1]:
        confusion_matrix = pd.crosstab(df_employment2[f[0]], df_employment2[f[1]])
        chitest = cramers_v(confusion_matrix.values)
        chi2.append(chitest)
    else:      # for same factor pair
        chi2.append(1)
```

In [ ]:
```python
chi2 = np.array(chi2).reshape((75,75))
```

In [ ]:
```python
chi2 = pd.DataFrame(chi2, index=df_employment2.columns.values, columns=df_employment2.columns.values) # then a
```

In [ ]:
```python
#chi2= p values
chi2.style.background_gradient(cmap='coolwarm')
```

- Above is a correlation table of our variables using a heat map style visualization to help identify levels of correlation between variables. In the above map, strongly correlated variables are colored red and uncorrelated or very weakly correlated variables are colored blue.

4.3 Data Quality Report

List the results of the data quality verification. If quality problems exist, suggest possible solutions. Solutions to data quality problems generally depend heavily on both data and business knowledge.

- Primary data quality issue is missing values in certain parts of the data. To correct for this, we imputed the mean value of the columns into those missing fields in order to have a more complete approximation of the missing data.

5. Stage Four - Modelling

As the first step in modelling, you'll select the actual modelling technique that you'll be using. Although you may have already selected a tool during the business understanding phase, at this stage you'll be selecting the specific modelling technique e.g. decision-tree building with C5.0, or neural network generation with back propagation. If multiple techniques are applied, perform this task separately for each technique.

5.1. Modelling technique

We opted for an unsupervised learning approach to our data, specifically clustering. We chose this method because with unsupervised learning, you can find relationships in data that aren't immmediately apparent. It does this by comparing the data based on similarity. In clustering, this is used to group the unlabeled data into clusters that represent patterns that were found in the data.

Our current focus is using K – Mode clustering, an extension of K-Means clustering, to determine the relationships that exist between our variables by presenting them as a group of visually clustered data points which reveal relationships not immediately apparent from the data set. The number of clusters is determined by the K value provided to the method, with several ways to determine the correct K value including an Elbow Chart or a Silhouette Graphs that can provide a visual representation of ideal K values for our data set.

We chose K-Mode clustering because of the nature of our data. K-Mode clusters are extremely efficient when working with large amounts of categorical data. K-Mode doesn't need to calculate all of the pair wise distances between data points. Additionally, it is distribution free, meaning that it does not require imposing distribution assumptions on the data in order to work.

K-Means clustering on the other hand does not work well with categorical data sets, primarily because it has the issues above that K-Mode does not, and ultimately it's objective function simply doesn't work with categorical data.

```
# First we will keep a copy of data
df_employment_copy = df_employment2.copy()
# Using K-Mode with "Cao" initialization
km_cao = KModes(n_clusters=2, init = "Cao", n_init = 1, verbose=1)
fitClusters_cao = km_cao.fit_predict(df_employment2)
```

```
# Predicted Clusters
fitClusters_cao
```

```
clusterCentroidsDf = pd.DataFrame(km_cao.cluster_centroids_)
clusterCentroidsDf.columns = df_employment2.columns
```

```
# Mode of the clusters
clusterCentroidsDf
```

```
# Using K-Mode with "Huang" initialization
km_huang = KModes(n_clusters=2, init = "Huang", n_init = 1, verbose=1)
fitClusters_huang = km_huang.fit_predict(df_employment2)
```

```
# Predicted clusters
fitClusters_huang
```

Building the clusters using K-Mode clustering.

- Clustering is a means of grouping data based on characteristics. The clusters will form based on relationships between the individual data points and potentially reveal new relationships that were not previously apparent.
- Clustering also allows for data compression when working with large data sets. After the data is clustered it can be referred to by cluster ID rather than the individual data points themselves.
- Another use of clustering is the preserve the privacy of the respondent. Data grouped in clusters is practically anonymized by its presence in the cluster and future reference by cluster ID. In our example, no individual respondent ID will need to be used going forward as we can refer to the groups by their cluster IDs.
- Here we used K-Mode clustering, which is similar to K - Means clustering but is based on the the Mode of K. K-Mode clusters are a good choice when working with large amounts of categorical data like what is present in our data set.

```
# Choosing K by comparing Cost against each K
cost = []
for num_clusters in list(range(1,10)):
    kmode = KModes(n_clusters=num_clusters, init = "Huang", n_init = 1, verbose=1)
    kmode.fit_predict(df_employment2)
    cost.append(kmode.cost_)
```

```
#For KModes, plot cost for a range of K values. Cost is the sum of all the dissimilarities between the clusters
# Select the K where you observe an elbow-like bend with a lesser cost value.
plt.subplots(figsize = (15,5))
y = np.array([i for i in range(1,10,1)])
plt.plot(y,cost)
```

- Here we used what is called the Elbow Method for determining the optimal value of 'k' which will also determine the number of clusters for our purposes. When examining the chart visually, one looks for the point at which the line makes an angle similar to that of an arm bent at the elbow. In our case there are a few bends, but the one with the least dispersion associated with it is at K=3, and as such that was the value we went forward with.

```
## Choosing K=3
km_cao = KModes(n_clusters=3, init = "Cao", n_init = 1, verbose=1)
fitClusters_cao = km_cao.fit_predict(df_employment2)
```

```
fitClusters_cao.shape
```

```
# Combining the predicted clusters with the original DF
df_employment2 = df_employment_copy.reset_index()
```

```
clustersDf = pd.DataFrame(fitClusters_cao)
clustersDf.columns = ['cluster_predicted']
combinedDf = pd.concat([df_employment2, clustersDf], axis = 1).reset_index()
combinedDf = combinedDf.drop(['index', 'level_0'], axis = 1)
```

```
combinedDf.head()
```

```
# Cluster Identification
cluster_0 = combinedDf[combinedDf['cluster_predicted'] == 0]
cluster_1 = combinedDf[combinedDf['cluster_predicted'] == 1]
cluster_2 = combinedDf[combinedDf['cluster_predicted'] == 2]
```

```
## Plotting the predicted cluster for first feature only
"""
i = 'TEERNPER_1.0'
plt.subplots(figsize = (15,5))
sns.countplot(x=combinedDf[i],order=combinedDf[i].value_counts().index,hue=combinedDf['cluster_predicted'])
plt.tight_layout()
plt.show()
"""
## Plotting the predicted cluster for all feature
for i in df_employment3.columns.values:
    if i == 'index':
        continue
    else:
        plt.subplots(figsize = (15,5))
        sns.countplot(x=combinedDf[i],order=combinedDf[i].value_counts().index,hue=combinedDf['cluster_predicte
        plt.tight_layout()
        plt.show()
    # comment out break statment to generate plots for all features
    break
```

- As the above plot shows, per predicted feature the survey repondent with hourly earnings are groupped in cluster 0 while cluster 1 contains the mix of respondent with an hourly earning & non hourly earnings.

5.4. Assess Model

The silhouette Method is also a method to find the optimal number of clusters and interpretation and validation of consistency within clusters of data. The silhouette method computes silhouette coefficients of each point that measure how much a point is similar to its own cluster compared to other clusters. by providing a succinct graphical representation of how well each object has been classified.

Model assessment - Summarise the results of this task, list the qualities of your generated models (e.g.in terms of accuracy) and rank their quality in relation to each other.

Revised parameter settings - According to the model assessment, revise parameter settings and tune them for the next modelling run. Iterate model building and assessment until you strongly believe that you have found the best model(s). Document all such revisions and assessments.

```
X = df_employment2
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_samples, silhouette_score

import matplotlib.pyplot as plt
import matplotlib.cm as cm
import numpy as np
#X = df_employment
range_n_clusters = [2, 3, 4, 5, 6, 7, 8]
for n_clusters in range_n_clusters:
    # Create a subplot with 1 row and 2 columns
    fig, (ax1) = plt.subplots(1)
    #fig, (ax1, ax2) = plt.subplots(1, 2)
    fig.set_size_inches(18, 7)

    # The 1st subplot is the silhouette plot
    # The silhouette coefficient can range from -1, 1 but in this example all
    # lie within [-0.1, 1]
    ax1.set_xlim([-0.1, 1])
    # The (n_clusters+1)*10 is for inserting blank space between silhouette
    # plots of individual clusters, to demarcate them clearly.
    ax1.set_ylim([0, len(X) + (n_clusters + 1) * 10])

    # Initialize the clusterer with n_clusters value and a random generator
    # seed of 10 for reproducibility.
    clusterer = KModes(n_clusters=n_clusters, random_state=10)
    #clusterer = KModes(n_clusters=num_clusters, init = "Cao", n_init = 1, verbose=1)
    cluster_labels = clusterer.fit_predict(X)

    # The silhouette_score gives the average value for all the samples.
    # This gives a perspective into the density and separation of the formed
    # clusters
    silhouette_avg = silhouette_score(X, cluster_labels)
    print("For n_clusters =", n_clusters,
          "The average silhouette_score is :", silhouette_avg)

    # Compute the silhouette scores for each sample
    sample_silhouette_values = silhouette_samples(X, cluster_labels)

    y_lower = 10
    for i in range(n_clusters):
        # Aggregate the silhouette scores for samples belonging to
        # cluster i, and sort them
        ith_cluster_silhouette_values = \
            sample_silhouette_values[cluster_labels == i]

        ith_cluster_silhouette_values.sort()

        size_cluster_i = ith_cluster_silhouette_values.shape[0]
        y_upper = y_lower + size_cluster_i

        color = cm.nipy_spectral(float(i) / n_clusters)
        ax1.fill_betweenx(np.arange(y_lower, y_upper),
                          0, ith_cluster_silhouette_values,
                          facecolor=color, edgecolor=color, alpha=0.7)

        # Label the silhouette plots with their cluster numbers at the middle
        ax1.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))

        # Compute the new y_lower for next plot
        y_lower = y_upper + 10  # 10 for the 0 samples

    ax1.set_title("The silhouette plot for the various clusters.")
    ax1.set_xlabel("The silhouette coefficient values")
    ax1.set_ylabel("Cluster label")

    # The vertical line for average silhouette score of all the values
    ax1.axvline(x=silhouette_avg, color="red", linestyle="--")

    ax1.set_yticks([])  # Clear the yaxis labels / ticks
    ax1.set_xticks([-0.1, 0, 0.2, 0.4, 0.6, 0.8, 1])
plt.show()
```

- Silhouette analysis can be used to study the separation distance between clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters.
- Observations from above Silhouette Plots:
- The silhouette plot shows that the n_cluster value of 3 is a bad pick, as all the points in the cluster with cluster_label=0 are below-average silhouette scores.
- The silhouette plot shows that the n_cluster value of 5 is a bad pick, as all the points in the cluster with cluster_label=2 and 4 are below-average silhouette scores.
- The silhouette plot shows that the n_cluster value of 6 is a bad pick, as all the points in the cluster with cluster_label=1,2,4 and 5 are below-average silhouette scores, and also due to the presence of outliers.
- Silhouette analysis is more ambivalent in deciding between 2 and 4.
- The thickness of the silhouette plot for the cluster with cluster_label=1 when n_clusters=2, is bigger in size owing to the grouping of the 3 sub-clusters into one big cluster.
- For n_clusters=4, all the plots are more or less of similar thickness and hence are of similar sizes, as can be considered as best 'k'.

```
#extract features and target variables
x = combinedDf.drop(columns="cluster_predicted")
y = combinedDf["cluster_predicted"]
#save the feature name and target variables
feature_names = x.columns
labels = y.unique()
#split the dataset
from sklearn.model_selection import train_test_split
X_train, test_x, y_train, test_lab = train_test_split(x,y,
                                                      test_size = 0.4,
                                                      random_state = 42)
```

- In the above cell, the data is being divided into multiple sub sets in order to both train and then subsequently test the our modeling solution against. This separation into multiple sets is in an effort to avoid the possibility of over fitting our model against the training set and thus skewing our results.

```
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier(max_depth =3, random_state = 42)
clf.fit(X_train, y_train)
```

```
#import relevant functions
from sklearn.tree import export_text
#export the decision rules
tree_rules = export_text(clf,
                        feature_names = list(feature_names))
#print the result
print(tree_rules)
```

- In the above cell we've created a decision tree to help determine which of our variables are associated with our other variables and how. Decision trees work sort of like flow charts of possible outcomes based on conditions that all start out at one root node. In this case the root of tree is the variable "TEI01ICD" and all of the branches of our tree extend from it.

```
test_pred_decision_tree = clf.predict(test_x)
```

```
#import the relevant packages
from sklearn import metrics
import seaborn as sns
import matplotlib.pyplot as plt
#get the confusion matrix
confusion_matrix = metrics.confusion_matrix(test_lab,
                                            test_pred_decision_tree)
#turn this into a dataframe
matrix_df = pd.DataFrame(confusion_matrix)
#plot the result
ax = plt.axes()
sns.set(font_scale=1.3)
plt.figure(figsize=(10,7))
sns.heatmap(matrix_df, annot=True, fmt="g", ax=ax, cmap="magma")
#set axis titles
ax.set_title('Confusion Matrix - Decision Tree')
ax.set_xlabel("Predicted label", fontsize =15)
#ax.set_xticklabels(['']+labels)
ax.set_ylabel("True Label", fontsize=15)
ax.set_yticklabels(list(labels), rotation = 0)
plt.show()
```

- The above visual is called a confusion matrix. These closesly resemble heatmaps and are used predominantly to help with statistical classification.
- Confusion matrices provide a way to visually observe the accuracy of a classification model based built from a know ground truth and branching into predicted values.

Assessing the accuracy of our completed model

```
metrics.accuracy_score(test_lab, test_pred_decision_tree)
```

```
print(metrics.classification_report(test_lab,
                                    test_pred_decision_tree))
```

```
#extract importance
importance = pd.DataFrame({'feature': X_train.columns,
'importance' : np.round(clf.feature_importances_, 3)})
importance.sort_values('importance', ascending=False, inplace = True)
print(importance)
```

```
from sklearn.model_selection import GridSearchCV
tuned_parameters = [{'max_depth': [1,2,3,4,5],
                     'min_samples_split': [2,4,6,8,10]}]
scores = ['recall']
for score in scores:

    print()
    print(f"Tuning hyperparameters for {score}")
    print()

    clf = GridSearchCV(
        DecisionTreeClassifier(), tuned_parameters,
        scoring = f'{score}_macro'
    )
    clf.fit(X_train, y_train)

    print("Best parameters set found on development set:")
    print()
    print(clf.best_params_)
    print()
    print("Grid scores on development set:")
    means = clf.cv_results_["mean_test_score"]
    stds = clf.cv_results_["std_test_score"]
    for mean, std, params in zip(means, stds,
                                 clf.cv_results_['params']):
        print(f"{mean:0.3f} (+/-{std*2:0.03f}) for {params}")
```

- The above indicates that our predictive model, as currently constructed has an accuracy of .97 or 97%. This indicates that our model is an excellent predictor of outcomes based on the relationships between our variables and which of them is likely to be indicative of another.

Model Ensemble

```
## Train Model
# Import the model we are using
from sklearn.ensemble import RandomForestClassifier
# Instantiate model with 10 decision trees
rf = RandomForestClassifier(n_estimators = 10, random_state = 42)
# Train the model on training data
rf.fit(X_train, y_train);
```

```
# Use the forest's predict method on the test data
predictions = rf.predict(test_x)
```

```
#Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(test_lab, predictions))
```

```
#### Finding Important Features in Scikit-learn
```

```
feature_list = list(X_train.columns)
feature_imp = pd.Series(rf.feature_importances_, index=feature_list).sort_values(ascending=False)
feature_imp_top20 = feature_imp.head(20)
feature_imp.head(20)
```

```
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
# Creating a bar plot
plt.figure(figsize=(10,7))
sns.barplot(x=feature_imp_top20, y=feature_imp_top20.index)
# Add labels to your graph
plt.xlabel('Feature Importance Score')
plt.ylabel('Features')
plt.title("Visualizing Important Features")
plt.legend()
plt.show()
```

Conclusion

Members of a sub class of the data set are best fits for targeting for advertisement of financial services products based on demographic data revealed in the survey used. The members of that sub class are easily identified by our algorithm for the purposes of identifying a target candidate pool to present to our customers. This suitability is based on a number of key factors such as employment and income level which indicate potential need or desire for specific financial services such as banks accounts or Certificates of Deposit (CDs).

Next Steps

- Business Application:

Now that we've built and assessed the accuracy of our model the next step is to apply it and interpret the results of our findings and how they apply to our ability to answer our questions. Based on the intermediate results of our model, it is clear that the clusters we have created are indicative of the predictive value of our variables. Going forward, we will determine the profile of the ideal candidate based on the output of our predictive modeling, identifying specifically which traits are most indicative of a candidate's suitability for our customers' services and be able to provide that information to clients for use in their marketing campaigns.

References

- Data Science in 5 minutes: What is one hot encoding? Educative. (n.d.). Retrieved February 9, 2022, from https://www.educative.io/blog/one-hot-encoding
- Decision trees in python - step-by-step implementation. AskPython. (2020, December 7). Retrieved February 10, 2022, from https://www.askpython.com/python/examples/decision-trees
- Franklin, S. J. (2019, November 26). Elbow method of k-means clustering algorithm. Medium. Retrieved February 9, 2022, from https://medium.com/analytics-vidhya/elbow-method-of-k-means-clustering-algorithm-a0c916adc540
- Google. (n.d.). Clustering algorithms | clustering in machine learning | google developers. Google. Retrieved February 10, 2022, from https://developers.google.com/machine-learning/clustering/clustering-algorithms
- Jaiswal, S. (2020, July 10). K-modesclustering. Medium. Retrieved February 10, 2022, from https://medium.com/@shailja.nitp2013/k-modesclustering-ef6d9ef06449
- Mean imputation for missing data (example in R & SPSS). Statistics Globe. (2022, January 19). Retrieved February 9, 2022, from https://statisticsglobe.com/mean-imputation-for-missing-data/
- Selecting the number of clusters with silhouette analysis on kmeans clustering. scikit. (n.d.). Retrieved February 9, 2022, from https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
- U.S. Bureau of Labor Statistics. (n.d.). Atus News releases. U.S. Bureau of Labor Statistics. Retrieved February 10, 2022, from https://www.bls.gov/tus/