Project 3

May 24, 2021

[288]:
```
import pandas as pd
import numpy as np
import scipy.stats as st
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
```

[228]:
```
df =pd.read_excel("Data-2.xlsx")
df
```

[228]:

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	T1	T2	T3	T4	T5
0	3	2	1	3	2	2003	1300.0	1	1	2	10	10	1	2	NaN	NaN	1.0
1	3	1	1	3	2	2003	1300.0	1	1	2	10	10	1	2	NaN	NaN	1.0
2	3	1	1	3	2	2003	1300.0	1	1	2	10	10	1	2	NaN	NaN	1.0
3	1	1	1	2	1	2013	900.0	2	1	4	10	8	2	1	NaN	NaN	2.0
4	1	2	1	2	1	2013	900.0	2	1	4	10	8	2	1	NaN	NaN	2.0
…	..	..	..	..	..	…	…	..	..	…	…	…	..	..	…	…	…
2230	2	2	1	3	2	2019	800.0	1	1	2	10	10	3	2	400.0	3000.0	NaN
2231	1	1	1	3	1	2019	800.0	1	1	2	9	9	1	3	500.0	3000.0	NaN
2232	5	1	1	3	1	2019	800.0	1	1	2	10	10	1	3	500.0	2500.0	NaN
2233	1	2	1	4	1	2019	1200.0	1	1	4	10	10	1	3	1000.0	4500.0	NaN
2234	2	1	1	4	2	2020	150.0	1	1	4	10	10	3	3	300.0	500.0	NaN

[2235 rows x 17 columns]

1 P1:

Compare the ratio of respondents who are concerned (category 1) and who are not concerned or not sure (categories 2 and 3) about the environmental issues among the two genders.

[229]:
```
# Ratio for men and women:
men = df[df['Q2'] == 2]
men_count = men.T1.value_counts().sum()
p_m_concerned = men.T1.value_counts()[1]/ men_count
p_m_concerned #  proportion of men concerned about the environment
```

[229]: 0.4789473684210526

[230]:
```
ci_plus = p_m_concerned + np.sqrt((0.5*0.5)/men_count)
ci_minus = p_m_concerned - np.sqrt((0.5*0.5)/men_count)
print(f'There is a 95 percent chance that true proportion of men that are␣
 ↪concerned about the environment lies between %.2f and %.2f'%(ci_minus,␣
 ↪ci_plus))
```

```
There is a 95 percent chance that true proportion of men that are concerned
about the environment lies between 0.46 and 0.49
```

[231]:
```
women = df[df['Q2'] == 1]
women_count = women.T1.value_counts().sum()
p_w_concerned = women.T1.value_counts()[1]/ women_count
p_w_concerned # proportion of women concerned about the environment
```

[231]: 0.6447488584474886

[232]:
```
ci_plus = p_w_concerned + np.sqrt((0.5*0.5)/women_count)
ci_minus = p_w_concerned - np.sqrt((0.5*0.5)/women_count)
print(f'There is a 95 percent chance that true proportion of men that are␣
 ↪concerned about the environment lies between %.2f and %.2f'%(ci_minus,␣
 ↪ci_plus))
```

```
There is a 95 percent chance that true proportion of men that are concerned
about the environment lies between 0.63 and 0.66
```

Apparently the two ratios are different between men and women but we will validate it in the next part

2 P2:

Find a statistical test method to see if the differences in P3 are statistically significant (note that you can use a method that was not presented in our course).

Using a two tailed Z -Test

[234]:
```
population_proportion = (men_count * p_m_concerned + women_count *␣
↪p_w_concerned) / (men_count + women_count)
population_proportion
```

[234]:
```
0.5601789709172259
```

[235]:
```
sdev_proportion_difference = np.sqrt(population_proportion␣
↪*(1-population_proportion) *((men_count + women_count)/(men_count *␣
↪women_count)))
sdev_proportion_difference
```

[235]:
```
0.021002968177941235
```

[236]:
```
z_obtained = (p_m_concerned - p_w_concerned)/sdev_proportion_difference
z_obtained
```

[236]:
```
-7.894193269338577
```

- Null Hypothesis: There is no difference in the proportion of people concerned for the environment between the two gender
- Alternate Hypothesis: There is significant difference between the proportion of people concerned for environment between the two genders

[192]:
```
z_critical = st.norm.ppf(0.975)
z_critical
```

[192]:
```
1.959963984540054
```

Since z obtained lies in critical region (i.e., it is less than -1.96) so we reject the null hypothesis and assert that there is a statistically significant difference between the proportion of females and males who are concerned about the environment.

3 P3:

Construct a random forest classifier model with T1 as the target variable and Q1 to Q5 as the independent variables. Sort the independent variables based on their feature importance in classifying T1. (hint: use this link (Links to an external site.))

[237]:
```
X = df[df.columns[:5]]
y = df['T1']
```

[238]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y,test_size =0.2)
```

[239]:
```
rfc = RandomForestClassifier()
rfc.fit(X_train, y_train)
rfc.feature_importances_
```

[239]:
```
array([0.19552212, 0.08168189, 0.40511677, 0.21859047, 0.09908874])
```

[240]:
```
rfc_imp_score = pd.DataFrame({"features" : df.columns[:5], "score" : rfc.
 ↪feature_importances_})
rfc_imp_score.sort_values("score", ascending = False, inplace =True)
```

[241]:
```
rfc_imp_score
```

[241]:

	features	score
2	Q3	0.405117
3	Q4	0.218590
0	Q1	0.195522
4	Q5	0.099089
1	Q2	0.081682

4 P4:

Repeat P1, but this time, compare T5 based on gender. Assume cold and moderate indoor temperatures as a category and hot indoor temperature as the second category.

[198]:
```
# T5 - Preferred Indoor temperature
    # Cold -1
    # Moderate -2
    # Hot - 3
#
```

[247]:
```
# Ratio for men and women:
men = df[df['Q2'] == 2]
men_count = men.T5.value_counts().sum()
p_m_preferred_temp = men.T5.value_counts()[3]/ men_count
women = df[df['Q2'] == 1]
women_count = women.T5.value_counts().sum()
p_w_preferred_temp = women.T5.value_counts()[3]/ women_count
population_proportion = (men_count * p_m_preferred_temp + women_count *␣
 ↪p_w_preferred_temp) / (men_count + women_count)
sdev_proportion_difference = np.sqrt(population_proportion␣
 ↪*(1-population_proportion) *((men_count + women_count)/(men_count *␣
 ↪women_count)))
z_obtained = (p_m_preferred_temp - p_w_preferred_temp)/
 ↪sdev_proportion_difference
z_obtained
```

[247]:
```
4.36583328592874
```

[248]:
```
z_critical = z_critical = st.norm.ppf(0.975)
z_critical
```

[248]: 1.959963984540054

- Null Hypothesis: There is no difference in the proportion of people who prefer hot indoor environment between the two gender
- Alternate Hypothesis: There is significant difference between the proportion of people who prefer hot indoor environment between the two genders

Using a two tailed Z -Test

Since Z obtained lies in critical region (i.e., z critical is equal to or great than 1.6) we reject the null hypothesis and assert that there is a statistically significant difference between the proportion of females and males who prefer hot indoor temperature

[ ]:

5 P5:

Recategorize Q11 and Q12 into two main groups instead of 10–> group 1 response: values 1 to 5 and group 2: response values 6 to 10. Then, repeat P1 to see if the ratio of indoor comfort levels and the importance of indoor comfort level are different in terms of the respondents' gender.

[250]:
```
bins = np.array([1,5,10])
df['Q11-1'] = pd.cut(df['Q11'], bins, labels= ['Group1', 'Group2'])
df['Q12-1'] = pd.cut(df['Q12'], bins, labels= ['Group1', 'Group2'])
```

[217]:
```
# Q11 Thermal comfort level
    # Group 1 - Poor Thermal Comfort Level
    # Group 2 - Good Thermal Comfort Level
```

[251]:
```
# Ratio for men and women:
men = df[df['Q2'] == 2]
men_count = men['Q11-1'].value_counts().sum()
p_m_thermal_comfort =men['Q11-1'].value_counts()[1]/ men_count
women = df[df['Q2'] == 1]
women_count = women['Q11-1'].value_counts().sum()
p_w_thermal_comfort = women['Q11-1'].value_counts()[1]/ women_count
```

[252]:
```
p_m_thermal_comfort # Sample proportion of men who have poor thermal comfort␣
 ↪level
```

[252]: 0.07901668129938542

[253]:
```
p_w_thermal_comfort # Sample proportion of women who have poor thermal comfort␣
 ↪level
```

[253]: 0.03290676416819013

[254]:
```
population_proportion = (men_count * p_m_thermal_comfort  + women_count *␣
 ↪p_w_thermal_comfort ) / (men_count + women_count)
sdev_proportion_difference = np.sqrt(population_proportion␣
 ↪*(1-population_proportion) *((men_count + women_count)/(men_count *␣
 ↪women_count)))
z_obtained = (p_m_thermal_comfort  - p_w_thermal_comfort )/
 ↪sdev_proportion_difference
z_obtained
```

[254]:
```
4.720540042297864
```

[256]:
```
z_critical = z_critical = st.norm.ppf(0.975)
z_critical
```

[256]:
```
1.959963984540054
```

- Null Hypothesis: There is no difference in the proportion of people who have poor thermal comfort between the two gender
- Alternate Hypothesis: There is significant difference between the proportion of people who have poor thermal comfort between the two gender

Using a two tailed Z -Test

Since Z obtained lies in critical region (i.e., z critical is equal to or great than 1.6) we reject the null hypothesis and assert that there is a statistically significant difference between the proportion of females and males who have poor thermal comfort

[ ]:
```
```

6 P6:

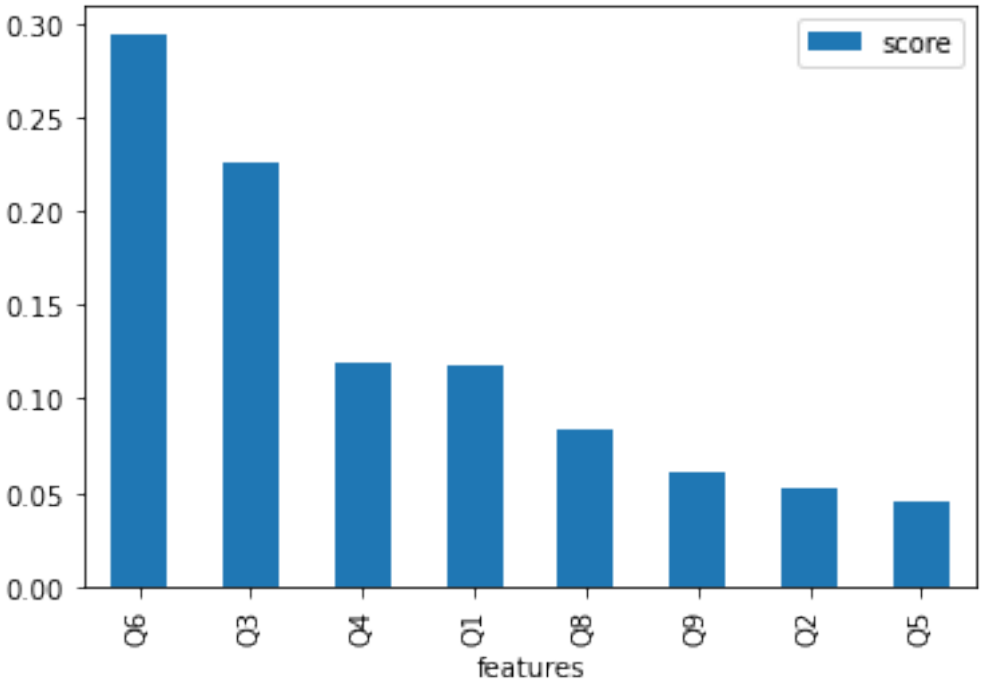
Repeat P3, but this time use T2 as the target variable and Q1 to Q9 (excluding Q7) as the independent variables (attributes). Can you name the main factors that determine a user's willingness in participating in DSM programs?

[268]:
```
X = df.loc[:,'Q1':'Q9'].drop('Q7',axis=1)
y= df['T2']
```

[269]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y,test_size =0.2)
rfc = RandomForestClassifier()
rfc.fit(X_train, y_train)
rfc.feature_importances_
rfc_imp_score = pd.DataFrame({"features" : X.columns, "score" : rfc.
 ↪feature_importances_})
rfc_imp_score.sort_values("score", ascending = False, inplace =True)
```

[272]:
```
rfc_imp_score.plot(kind ='bar',x= 'features', y= 'score')
```

[272]: <matplotlib.axes._subplots.AxesSubplot at 0x7f0a52dc0f70>



The four most important features that determine a user's willingness in participating in Demand Side Management are:

- House Construction Year
- Respondent's Race
- Household Monthly Income
- Respondent's Age

[ ]:
```

```

7 P7:

Find the average of T3 and T4, drop the null values, and normalize the average based on the floor area of houses. Create a random forest regressor, and sort the feature importance based on the normalized value as the target and Q6, Q8, Q9, and Q10. What are the most important features impacting monthly consumption

[277]:
```
df_c = df.dropna(subset=['T3','T4'])
```

[ ]:
```
df_c['Average Bill'] = (df_c.T3 +df_c.T4)/2
```

[ ]:
```
df_c['Bill/Area'] = df_c['Average Bill'] / df_c.Q7
```

[285]:
```
X = df_c[['Q6','Q8','Q9','Q10']]
y = df_c['Bill/Area']
```

[289]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y,test_size =0.2)
rfr = RandomForestRegressor()
rfr.fit(X_train, y_train)
rfr.feature_importances_
rfr_imp_score = pd.DataFrame({"features" : ["Vintage", "Ownership","Modern␣
↪Structure","Air-Conditioning Technology"], "score" : rfr.
↪feature_importances_})
rfr_imp_score.sort_values("score", ascending = False, inplace =True)
```

[290]:
```
rfr_imp_score
```

[290]:

	features	score
0	Vintage	0.488056
1	Ownership	0.315622
3	Air-Conditioning Technology	0.153969
2	Modern Structure	0.042353

The most important features impacting monthly consumption are House Vintage, Ownership, and the kind of air conditioning technology installed

[ ]:
```

```