part_3_segementation_of_orange_var

January 24, 2025

1 Introduction

In this part of the analysis, we focused on reassigning individuals into the identified clusters using illustrative variables. Building on the clustering results from the first step, we evaluated two algorithms, Random Forest and Gradient Boosting. Here, we focus on the segmentation of the orange variables. In the first part, using the green variables and then using a specific set of variables.

[1]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import SpectralClustering
import matplotlib.pyplot as plt
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.decomposition import PCA
from sklearn.utils.class_weight import compute_class_weight
import numpy as np
from sklearn.model_selection import GridSearchCV
```

1.1 Data Loading

[2]:
```
file_path = "fic_epita_kantar_codes.csv"
data = pd.read_csv(file_path, sep=';')

# Define Orange and Green variable groups
orange_vars = [
    "A9_1_slice", "A9_2_slice", "A9_3_slice", "A9_4_slice", "A9_5_slice",
    "A9_6_slice", "A9_7_slice", "A9_8_slice", "A9_9_slice", "A9_10_slice",
    "A9_11_slice", "A9_12_slice", "A9_13_slice", "A9_14_slice", "A9_15_slice",
    "A9_16_slice", "A10_1_slice", "A10_2_slice", "A10_3_slice", "A10_4_slice",
    "A10_5_slice", "A10_6_slice", "A10_7_slice", "A10_8_slice",
    "A11_1_slice", "A11_2_slice", "A11_3_slice", "A11_4_slice", "A11_5_slice",
    "A11_6_slice", "A11_7_slice", "A11_8_slice", "A11_9_slice", "A11_10_slice",
    "A11_11_slice", "A11_12_slice", "A11_13_slice"
]
green_vars = [
```

```
    "A11_1_slice", "A12", "A13", "A14", "A4", "A5", "A5bis", "A8_1_slice",
    "A8_2_slice", "A8_3_slice", "A8_4_slice", "B1_1_slice", "B1_2_slice",
    "B2_1_slice", "B2_2_slice", "B3", "B4", "B6", "C1_1_slice", "C1_2_slice",
    "C1_3_slice", "C1_4_slice", "C1_5_slice", "C1_6_slice", "C1_7_slice",
    "C1_8_slice", "C1_9_slice"
]
specific_vars = [
    "rs3", "rs5", "rs6", "RS1", "RS191", "RS192", "RS193", "RS102RECAP",
    "rs11recap2", "RS11recap", "RS193bis", "RS2Recap", "RS56Recap",
    "RS2", "RS11", "RS102"
]
```

1.2 Adding the cluster column using the best algo with the best number of cluster found in part 1

[3]:
```
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

# Fill missing values in Orange variables
orange_data = data[orange_vars].fillna(0)

# Scale the data
scaler = StandardScaler()
scaled_orange_data = scaler.fit_transform(orange_data)

# Perform KMeans clustering with 3 clusters
kmeans = KMeans(n_clusters=3, random_state=42)
data['cluster_orange'] = kmeans.fit_predict(scaled_orange_data)
```

1.3 Segmentation based on green variable

[4]:
```
# Extract features (Green variables) and target (Orange clusters)
X_green = data[green_vars].fillna(0)
y_orange = data['cluster_orange']

# Split data into training and test sets
X_train_green, X_test_green, y_train, y_test = train_test_split(
    X_green, y_orange, test_size=0.2, random_state=42
)

# Train and Evaluate model
rf_green = RandomForestClassifier(random_state=42)
rf_green.fit(X_train_green, y_train)

y_pred_green = rf_green.predict(X_test_green)
print("Classification Report (Green Variables) with Random Forest:\n",␣
 ↪classification_report(y_test, y_pred_green))
```

```
Classification Report (Green Variables) with Random Forest:
              precision    recall  f1-score   support

           0       0.77      0.71      0.74       262
           1       0.74      0.56      0.64       184
           2       0.75      0.84      0.79       554

    accuracy                           0.75      1000
   macro avg       0.75      0.70      0.72      1000
weighted avg       0.75      0.75      0.75      1000
```

[5]:
```
# Extract features (Green variables) and target (Orange clusters)
X2 = data[green_vars].fillna(0)
y2 =  data['cluster_orange']

# Split data into training and test sets
X_train2, X_test2, y_train2, y_test2 = train_test_split(
    X2, y2, test_size=0.2, random_state=42
)

# Train and Evaluate model
gb = GradientBoostingClassifier(random_state=42)
gb.fit(X_train2, y_train2)
y_pred2 = gb.predict(X_test2)
print("Classification Report (Specific Variables with Gradient Boosting):\n",␣
 ↪classification_report(y_test2, y_pred2))
```

```
Classification Report (Specific Variables with Gradient Boosting):
              precision    recall  f1-score   support

           0       0.75      0.68      0.71       262
           1       0.73      0.62      0.67       184
           2       0.75      0.82      0.79       554

    accuracy                           0.75      1000
   macro avg       0.75      0.71      0.73      1000
weighted avg       0.75      0.75      0.75      1000
```

1.4 Segmentation based on specific variable

[6]:
```
# For Random Forest
# Extract features (Specific variables) and target (Orange clusters)
X_specific = data[specific_vars].fillna(0)
y_orange = data['cluster_orange']

# Split data into training and test sets
```

```
X_train_specific, X_test_specific, y_train, y_test = train_test_split(
    X_specific, y_orange, test_size=0.2, random_state=42
)

# Compute class weights
class_weights = compute_class_weight(
    class_weight='balanced',
    classes=np.unique(y_train),
    y=y_train
)

# Convert weights into a dictionary
class_weights_dict = {i: weight for i, weight in enumerate(class_weights)}

# Train and Evaluate model
rf_specific = RandomForestClassifier(random_state=42,␣
 ↪class_weight=class_weights_dict)
rf_specific.fit(X_train_specific, y_train)
y_pred_specific = rf_specific.predict(X_test_specific)
print("Classification Report (Specific Variables) with Random Forest:\n",␣
 ↪classification_report(y_test, y_pred_specific))
```

```
Classification Report (Specific Variables) with Random Forest:
              precision    recall  f1-score   support

           0       0.32      0.32      0.32       262
           1       0.21      0.20      0.20       184
           2       0.55      0.55      0.55       554

    accuracy                           0.43      1000
   macro avg       0.36      0.36      0.36      1000
weighted avg       0.43      0.43      0.43      1000
```

[7]:
```
# For Gradient Boosting
# Extract features (Specific variables) and target (Orange clusters)
X_specific = data[specific_vars].fillna(0)

# Split data into training and test sets
X_train_specific, X_test_specific, y_train, y_test = train_test_split(
    X_specific, y_orange, test_size=0.2, random_state=42
)

# Train and Evaluate model
gb_specific = GradientBoostingClassifier(random_state=42)
gb_specific.fit(X_train_specific, y_train)
y_pred_specific = gb_specific.predict(X_test_specific)
```

```
print("Classification Report (Specific Variables) with Gradient Boosting:\n",␣
 ↪classification_report(y_test, y_pred_specific))
```

```
Classification Report (Specific Variables) with Gradient Boosting:
              precision    recall  f1-score   support

           0       0.32      0.05      0.09       262
           1       0.34      0.07      0.12       184
           2       0.55      0.91      0.69       554

    accuracy                           0.53      1000
   macro avg       0.40      0.35      0.30      1000
weighted avg       0.45      0.53      0.43      1000
```

1.5 Optimizing results to try to improve the average performance we get

[8]:
```
# For Random Forest
# Extract features (Specific variables) and target (Orange clusters)
X_specific = data[specific_vars].fillna(0)

# Split data into training and test sets
X_train_specific, X_test_specific, y_train, y_test = train_test_split(
    X_specific, y_orange, test_size=0.2, random_state=42
)

# Define parameter grid for Random Forest
param_grid = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'bootstrap': [True, False]
}

# Initialize a Random Forest classifier
rf = RandomForestClassifier(random_state=42)

# Perform GridSearchCV to optimize hyperparameters
grid_search = GridSearchCV(estimator=rf, param_grid=param_grid,
                           scoring='accuracy', cv=3, verbose=0, n_jobs=-1)

# Fit the GridSearchCV on training data
grid_search.fit(X_train_specific, y_train)

# Get the best parameters and score
best_params = grid_search.best_params_
```

```
best_score = grid_search.best_score_
print(f"Best Hyperparameters: {best_params}")
print(f"Best Cross-Validation Accuracy: {best_score:.4f}")

# Train and Evaluate model with optimized hyperparameters
rf_optimized = RandomForestClassifier(**best_params, random_state=42)
rf_optimized.fit(X_train_specific, y_train)
y_pred_specific_optimized = rf_optimized.predict(X_test_specific)
print("Classification Report (Specific Variables with Hyperparameter␣
 ↪Optimization):\n",
      classification_report(y_test, y_pred_specific_optimized))
```

```
Best Hyperparameters: {'bootstrap': True, 'max_depth': 10, 'min_samples_leaf':
2, 'min_samples_split': 10, 'n_estimators': 200}
Best Cross-Validation Accuracy: 0.5153
Classification Report (Specific Variables with Hyperparameter Optimization):
              precision    recall  f1-score   support

           0       0.20      0.03      0.06       262
           1       0.32      0.08      0.12       184
           2       0.55      0.90      0.68       554

    accuracy                           0.52      1000
   macro avg       0.36      0.34      0.29      1000
weighted avg       0.42      0.52      0.42      1000
```

[9]:
```
# For Gradient Boosting
# Define parameter grid for Gradient Boosting
param_grid2 = {
    'n_estimators': [50, 100, 200],
    'learning_rate': [0.01, 0.1, 0.2],
    'max_depth': [3, 5, 10],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}

# Initialize a Gradient Boosting classifier
gb2 = GradientBoostingClassifier(random_state=42)

# Perform GridSearchCV to optimize hyperparameters
grid_search2 = GridSearchCV(estimator=gb2, param_grid=param_grid2,
                            scoring='accuracy', cv=3, verbose=0, n_jobs=-1)

# Fit the GridSearchCV on training data
grid_search2.fit(X_train2, y_train2)
```

```python
# Get the best parameters and score
best_params2 = grid_search2.best_params_
best_score2 = grid_search2.best_score_
print(f"Best Hyperparameters (Gradient Boosting): {best_params2}")
print(f"Best Cross-Validation Accuracy (Gradient Boosting): {best_score2:.4f}")

# Train and Evaluate model with optimized hyperparameters
gb_optimized = GradientBoostingClassifier(**best_params2, random_state=42)
gb_optimized.fit(X_train2, y_train2)
y_pred2_optimized = gb_optimized.predict(X_test2)
print("Classification Report (Specific Variables with Gradient Boosting␣
 ↪Optimization):\n",
      classification_report(y_test2, y_pred2_optimized))
```

```
Best Hyperparameters (Gradient Boosting): {'learning_rate': 0.1, 'max_depth': 3,
'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 50}
Best Cross-Validation Accuracy (Gradient Boosting): 0.7290
Classification Report (Specific Variables with Gradient Boosting Optimization):
              precision    recall  f1-score   support

           0       0.74      0.69      0.72       262
           1       0.75      0.59      0.66       184
           2       0.75      0.83      0.79       554

    accuracy                           0.75      1000
   macro avg       0.75      0.70      0.72      1000
weighted avg       0.75      0.75      0.75      1000
```

1.5.1 Conclusion

In this part of the analysis, we successfully reassigned individuals to their respective clusters using the illustrative variables. The segmentation achieved good performance when using the Green variables, demonstrating their strong predictive power for reassignment tasks. However, performance was average when using the specific given variables. We have however pretty good result after optimizing hyperparameters for the Gradient Boostin with an accuracy of aroun 0.75.

Between the two algorithms tested, Gradient Boosting showed better results compared to Random Forest, reinforcing its ability to handle complex, non-linear relationships in the data. Overall, Gradient Boosting is recommended for its superior performance, especially when working with the specific given variables.