PERFORMANCE ASSESSMENT

Task 2 | Dimensionality Reduction

KAILI HAMILTON

Masters of Science in Data Analytics, Western Governors University

Course: D212

Instructor: Dr. Kesselly Kamara

Program Mentor: Krissy Bryant

February, 2024

TABLE OF CONTENTS

A1

One research question for this analysis is, what are the principal features of the telecommunications dataset and how can we use those to support business decisions?

A2

One goal of this data analysis is to reduce the dimensions of this dataset to its principal components by keeping features with significant variance while minimizing information loss.

B1

Principal Component Analysis (PCA) is a dimensionality reduction technique for machine learning. The expected outcomes of PCA are to reduce the number of dimensions (i.e., features or variables) to its most important features thereby reducing the noise of irrelevant features in order to make effective predictions.

PCA is used on continuous variables that have been Normalized. PCA identifies the principal components by keeping features with significant variance while minimizing information loss. This is accomplished by "decorrelating" the variables, meaning the original variables are transformed by rotating the data to be aligned with the axes and shifts the data so they have a mean of 0. When the data has been fit and transformed as described on all components, we can identify the principal components by creating a scree plot and using the elbow method. Principal components are those that are the directions in which the data varies the most, i.e., highest variance. Now, refit PCA with the number of components identified and use those principal components for further analysis (Wilson, n.d.).

B2

PCA assumes that features of high variance are informative and thus discards features of low variance. If there isn't high collinearity amongst variables, PCA may not be the right dimension reduction technique.

C1

The continuous variables that will be used to answer the PCA question from Part A are as follows:

Lat	Lng	Population	Age	Income	Outage_sec_perweek	Email
Contacts	Yearly_equip_failure		Tenure	MonthlyCharge	Bandwidth_GB_Year	

C2

The dataset was standardized using RobustScaler from the scikit learn library. A screenshot is included below. A copy of the cleaned dataset is included with the submission of this performance assessment.

Data Normalization

```
robust = RobustScaler()
scaled_data = robust.fit_transform(df)
scaled_data
```

```
array([[ 2.49150047, -2.67486933, -0.23109413, ..., -0.53461297,
         0.08181679, -0.54601855],
       [ 0.72920498,  0.21642847,  0.60623492, ..., -0.63988949,
         1.23689002, -0.56982572],
       [ 0.88100806, -2.07882957,  0.06633146, ..., -0.36735607,
        -0.12405657, -0.28159143],
       ...,
       [-0.57285501, -0.73690422, -0.20148833, ...,  0.22378478,
         0.04097408,  0.20226104],
       [-0.85965576,  0.16396251,  2.62787611, ...,  0.66586433,
         1.40134341,  0.73314055],
       [-0.69296594,  0.25787352,  0.74975865, ...,  0.52127068,
         0.82295943,  0.59269989]])
```

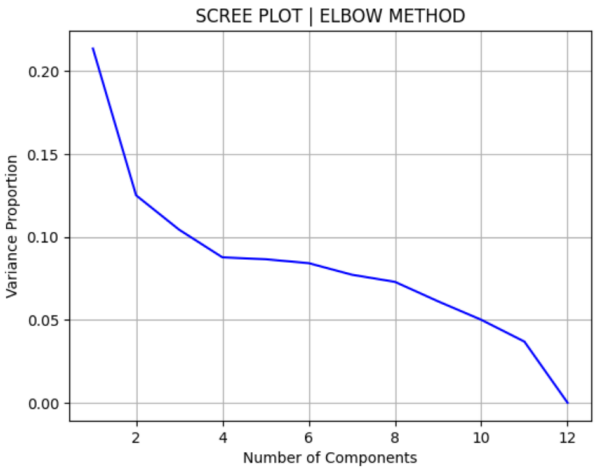
D1

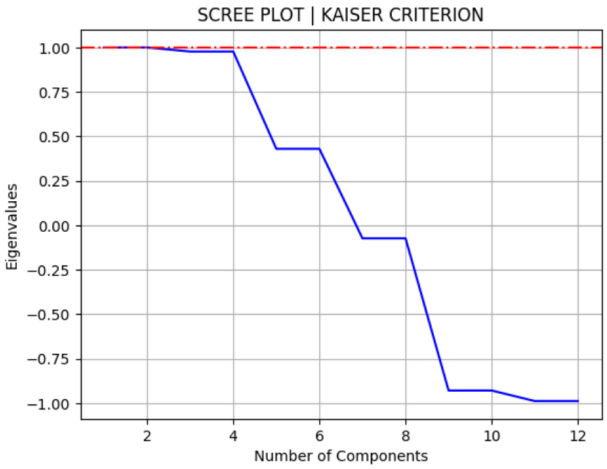
The loading matrix of all the principal components is presented in a screenshot here:

	Lat	Lng	Population	Age	Income	Outage_sec_perweek	Email	Contacts	Yearly_equip_failure	Tenure	MonthlyCharge	Bandwidth_GB_Year
PC1	-0.257290	-0.048289	0.964702	0.006363	-0.012815	0.004480	0.023713	0.002020	-0.002291	-0.000641	-0.004664	-0.000819
PC2	0.335223	-0.939236	0.043795	-0.012438	0.025845	-0.011272	-0.041878	-0.006494	0.004650	-0.010933	-0.022802	-0.011723
PC3	-0.005011	0.029332	0.014818	-0.004242	0.996728	-0.051919	-0.048642	0.000575	0.010996	0.006704	-0.011307	0.007412
PC4	-0.471114	-0.207835	-0.155309	-0.003735	0.038165	-0.160069	0.825117	-0.002072	-0.021890	-0.023423	-0.034310	-0.024792
PC5	-0.189695	-0.043080	-0.042485	0.008944	-0.063610	-0.915859	-0.307573	-0.017844	0.014994	0.005525	-0.149180	-0.001094
PC6	-0.736299	-0.260517	-0.199055	-0.012782	-0.000533	0.314970	-0.449596	0.002143	0.059498	0.138641	0.070710	0.145258
PC7	0.099206	0.005696	0.029580	0.023429	-0.002112	-0.174296	0.086674	0.003829	-0.005083	0.403791	0.768164	0.444570
PC8	-0.103304	-0.035388	-0.024830	0.012750	0.009331	-0.049711	-0.090597	-0.000847	-0.064410	-0.554806	0.614428	-0.535770
PC9	0.027979	0.012730	0.009139	0.046818	-0.008472	-0.011105	0.044915	-0.010438	0.994633	-0.042188	0.040100	-0.042140
PC10	0.006036	0.015592	0.009203	-0.996879	-0.005605	-0.015636	0.003927	-0.050211	0.044466	-0.015157	0.025529	0.014051
PC11	0.000045	-0.005804	-0.001906	-0.049622	-0.001791	-0.017704	-0.002854	0.998489	0.012722	-0.003905	-0.003759	-0.001999
PC12	0.000663	0.000268	0.000014	0.019915	-0.000691	0.000147	0.000109	-0.000507	-0.000188	-0.712147	-0.032038	0.701015

D2

The total number of principal components using the elbow rule from a scree plot and using the Kaiser criterion is two. The scree plot is included below.





D3

The variance for each of the two principal components are 0.213 and 0.125.

```
array([0.21334604, 0.12503905])
```

D4

The total variance captured by the principal components identified in part D2 is about 34%

```
exp_var_2 = pca_2.explained_variance_ratio_
exp_var_2
```

```
array([0.21334604, 0.12503905])
```

```
total_variance = np.sum(exp_var_2)
```

```
total_variance
```

```
0.33838508215174223
```

D5

In summary, the results of this PCA are that the features, i.e., dimensions, recorded about the customers in the telecommunications dataset can be reduced to two principal components while minimizing data loss. These components can explain 34% of the variance in the dataset and can be used for future machine learning analysis.

Furthermore, the loadings matrix indicates that Principal Component 0 is highly, positively, correlated with “Population”, Principal Component 1 is highly negatively correlated with “Longitude”.

```
loadings_2 = pca_2.components_
```

```
loadings_2_df = pd.DataFrame(loadings_2, columns=scaled_df.columns)
```

```
print("Loadings of features on each principal component: ")
loadings_2_df
```

```
Loadings of features on each principal component:
```

	Lat	Lng	Population	Age	Income	Outage_sec_perweek	Email	Contacts	Yearly_equip_failure	Tenure	MonthlyCharge	Bandwidth_GB_Year
0	-0.257290	-0.048289	0.964702	0.006363	-0.012815	0.004480	0.023713	0.002020	-0.002291	-0.000641	-0.004664	-0.000819
1	0.335223	-0.939236	0.043795	-0.012438	0.025845	-0.011272	-0.041878	-0.006494	0.004650	-0.010933	-0.022802	-0.011723

```
loadings_2_df.max(axis=1)
```

```
0    0.964702
1    0.335223
dtype: float64
```

```
loadings_2_df.min(axis=1)
```

```
0   -0.257290
1   -0.939236
dtype: float64
```

E

Kamara, Kesselly. “D212 Recommended Study Plan”. D212 Western Governors University, n.d., srm.file.force.com/servlet/fileField?id=0BE3x000000c9YS.

Wilson, Benjamin. “Unsupervised Learning in Python”. n.d., app.datacamp.com/learn/courses/unsupervised-learning-in-python

F

Kamara, Kesselly. “D212 Recommended Study Plan”. D212 Western Governors University, n.d., srm.file.force.com/servlet/fileField?id=0BE3x000000c9YS.

Wilson, Benjamin. “Unsupervised Learning in Python”. n.d., app.datacamp.com/learn/courses/unsupervised-learning-in-python

G

Professional communication is demonstrated throughout the content and presentation of this PA.