Churn Analysis Using Principal Component Analysis (PCA)

Douglas Ehlert

Churn Analysis Using Principal Component Analysis (PCA)

Part 1: Research Question

To maintain profits, it is critical for a telecommunication organization to retain customers for as long as possible. This helps to offset customer acquisition costs and reduces the dollars spent on sales and marketing. This report will examine the effects of numerous variables on customer churn. The research question is: it possible to classify (or predict) whether a customer will churn using an unsupervised learning technique such as PCA and logistic regression? The goal of the analysis is to enable the organization to predict customer churn.

Part II: Technique Justification

PCA utilizes unsupervised learning to categorize data. Often utilized with data sets that have many dimensions, PCA reduces those dimensions to reduce overfitting and can help make the model simpler. PCA accomplishes this through feature extraction that leads to new dimensions, that will take a large amount of variance into consideration, and by dropping other dimensions that don't have a high level of variance. To summarize, PCA "is the process of computing transformed variables or new dimension along the direction of major variability in the data" (Sharma, 2021).

This project will use PCA to analyze the supplied customer churn data set. By measuring data variance and using feature extraction, the algorithm will create new dimensions, that account for a large amount of variance, and will decrease the number of dimensions used in the Logistic Regression that will predict customer churn. Finally, the analysis will provide an accuracy score for the Logistic Regression model.

An assumption made by PCA is that the analysis includes multiple variables, measured continuously (Bruce et al., 2020). This will be assured in the selection of variables. A list of imported libraries and packages follows:

- Pandas
    - Used for data manipulation such as import and of export of .csv files, data wrangling, analysis, and cleaning
- MatPlotLib
    - Used for data visualization (ROC-AUC Curve)
- Numpy
    - Used for mathematical operatons
- StandardScaler
    - Used to normalize the data
- PCA
    - Used to instantiate the model for analysis
- Pipeline
    - Used to create a pipeline to standardize and run PCA in one step
- LogisticRegression
    - Used to create a Logistic Regression model and predict accuracy
- Train_test_split
    - Used to split the data into training and testing sets for Logistic Regression

Part III: Data Preparation

The initial data preparation goal is to import the dataset and libraries/packages, perform an overview of the data (columns, datatype, etc.). Categorical variables were then dropped

because PCA works best with continuous variables. 'Case Order', 'Customer_id', 'Interaction', 'UID' were dropped because they are simply identifiers and were not pertinent to this analysis. The data was then checked for any null values and exported as 'data_clean.csv'. The code for the above description is contained in cells 1-7 of the Jupyter Notebook included in the submission. A heat map of correlation was run, in cell eight of the attached Jupyter Notebook. It shows a strong correlation between "Tenure" and "Bandwidth_GB_Year".

One data preparation technique used in this analysis is normalization of the data. The data must be normalized for PCA. For this analysis, this is completed in a pipeline and shown in cell 9 of the attached Jupyter Notebook.

All the variables, used in this analysis, are continuous and include Population, Children, Age, Income, Outage_sec_perweek, email, contacts, Yearly_equip_failure, Tenure, MonthlyCharge, and Bandwidth_GB_Year.

Part IV: Analysis

Cell nine in the Jupyter Notebook shows the PCA instance being initialized and run in a pipeline, with the data also being standardized in the same step.

The principal components (matrix) and variances can be found in the output of cells 10 and 11 respectively. Additionally, cumulative variance can be found in cell 12. If all components are used, there are 11 in total, the cumulative variance adds to 100, which is expected in a PCA model.

Two principal components were selected based on the elbow method and the scree plot shown in cell 13 of the Jupyter Notebook. Using two principal components will result in 27.74

total variance explained. Although this is not ideal, the scree plot shows that increasing the number of components will lead to only marginal gains in model effectiveness. This is deduced from the flat line that extends from component two to component nine in the scree plot and can be visually confirmed in the output of cell 13. Cells 14 and 15 of the attached Jupyter Notebook visualize the churn variable in relation to the two principal components that were selected for this visualization. One can see that the positive churn data points, shown as orange dots in the cell 15 output of the Jupyter Notebook, are congregated in the left cluster. This is a visual demonstration of the effectiveness of this PCA analysis.

Cells 16 and 17 of the attached Jupyter Notebook, are confirmation of the variances that were found in earlier steps.

Cell 18 splits the data in a train and test set and applies numerical values to the 'True' or 'False' that is currently found in the Churn data column.

Cell 19 runs the PCA model as a logistic regression. The output of cell 19 shows the accuracy of the model with various numbers of components and is copied below for reference.

```
2 0.75
3 0.7748
4 0.7872
5 0.7888
6 0.8048
7 0.8072
8 0.81
9 0.81
10 0.836
11 0.8512
```

The above values show that using only two PCA components will result in a Logistic Regression model accuracy of .75. Increasing to all 11 components will increase that accuracy to .85. It would be recommended to run the PCA model with all 11 components if the compute power is available.

Using the above Logistic Regression model, based on PCA, the organization can predict whether a customer will churn with an accuracy of 85%. If a customer is predicted to churn, they could be placed in a specialized “Customer Retention” pool that involves more in-depth customer service in an effort to prevent the customer from churning.

References

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical Statistics for Data Scientists*. O'Reilly Media, Inc.

Sharma, A. (2021, May 3). *PCA or principal component analysis on Customer Churn Data*. Medium. Retrieved May 31, 2022, from https://medium.com/data-science-on-customer-churn-data/pca-or-principal-component-analysis-on-customer-churn-data-d18ca60397ed#:~:text=The%20analysis%20shows%20that%20PCA,first%20and%20the%20last%20components).