Analysis Steps and Results

1. The goal of the analysis is to identify the characteristics of customers who are leaving to help the company reduce the churn rate. The techniques used are Factor Analysis of Mixed Data (FAMD) as a descriptive method and Logistic Regression as a classification method.
2. Factor Analysis of Mixed Data (FAMD) is an unsupervised learning method. The goal of unsupervised methods is to discover useful things about the data, such as subgroups among the variables or the observations. We are not trying to make a prediction when applying an unsupervised learning method. In fact, we don’t have a response variable, like we do, for example, in a regression analysis. Unsupervised methods are often used for exploratory data analysis. It’s a good tool to explore and visualize the data before supervised learning methods are applied.

One of the unsupervised methods is Principal Component Analysis (PCA). It is commonly used to summarize a large set of numeric variables and identify a few important ones that explain most of the variance in the data. Another unsupervised method is Clustering. It is used to segment the data into groups with some similarities.

Since the given dataset contained variables of mixed type (16 qualitative and 3 quantitative variables), we could not use PCA or Clustering. Instead, we used Factor Analysis of Mixed Data (FAMD). FAMD can help us reduce the number of variables and select the most relevant ones.

1. Graphical analysis shows that customers on a month-to-month contract have the highest churn rate.

Chart, bar chart

Description automatically generated

The same is shown with the cross-tabulation table (Contract vs. Churn). Most customers who churn are on a month-to-month contract. Hardly anyone with a two-year contract has left. Those who stay are almost evenly split between three types of contract length.

Graphical user interface, text

Description automatically generated

We ran a chi-squared test to see if the association between Contract and Churn is statistically significant.

Text

Description automatically generated with low confidence

Since the p-value is less than 0.05, the correlation is statistically significant. Therefore, we conclude that there is a relationship between Contract and Churn.

1. **Factor Analysis of Mixed Data (FAMD).**

We calculated the proportion of variance explained by each dimension and visualized it with a scree plot.

Text

Description automatically generated

Chart, histogram

Description automatically generated

Typically, we use one as a cutoff point and retain dimensions with an eigenvalue of more than one for further analysis. The first two dimensions have eigenvalues 5.256474 and 2.856002, so we will focus on them.

Ideally, we are looking for a reduced set of variables to explain 80% or more of the total variance. The cumulative variance of all five dimensions is 52.99%, with the first two dimensions only explaining 35.27% of the total variance. The relationship among the variables may be non-linear. Plus, some informative variables might be missing. Therefore, reducing this dataset to a few critical variables will not be straightforward.

Let’s see what variables the principal dimensions consist of. The following output shows that Dimension 1 is high on MonthlyCharges and Total Charges. Dimension 2 is high on Contract, InternetService, PaymentMethod, and Tenure.

Dim.1 Dim.2 Dim.3 Dim.4 Dim.5

Tenure 7.978791 **9.921098** 3.204640 1.991816 4.415041

MonthlyCharges **13.562857** 7.005692 0.134895 0.093491 3.587336

TotalCharges **15.166640** 1.423075 1.654367 1.208909 0.028437

Gender 0.004310 0.005650 0.000079 0.112023 0.051483

SeniorCitizen 0.583880 4.835316 0.309718 0.521700 32.198612

Partner 1.820855 5.679346 3.649290 32.364337 2.136338

Dependents 0.018297 8.236209 1.425984 41.955883 2.175431

PhoneService 0.063721 1.062419 **26.087403** 2.089533 15.781726

MultipleLines 5.039606 0.801185 10.234627 0.830706 0.009699

InternetService 8.015633 **14.620604** **34.052579** 1.300913 0.987998

OnlineSecurity 4.096996 2.994373 5.653308 0.652720 2.162163

OnlineBackup 6.422570 0.181664 1.464935 0.135413 0.000476

DeviceProtection 7.659598 0.215437 1.715855 0.098776 0.272648

TechSupport 5.093831 2.625844 6.236395 1.035643 3.686491

StreamingTV 8.512198 1.009345 0.326744 1.008007 0.628388

StreamingMovies 8.593364 0.976084 0.450485 0.394154 0.302054

Contract 1.989980 **19.173755** 2.581503 3.002218 12.088817

PaperlessBilling 1.650735 6.497506 0.055634 2.471385 2.512643

PaymentMethod 3.726139 **12.735397** 0.761560 8.732371 16.974220

We can also visualize the contributions to dimensions with bar graphs.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

In general, any variable that goes above the red dashed line contributes more than expected on average to the dimension and can be included in a predictive model.

The graph shows that MonthlyCharges, TotalCharges, Contract, and InternetService will be the first candidates to be included in the predictive analysis.

Lastly, we will visualize the contributions using a Squared Loading Plot. The Squared Loading plot shows both quantitative and qualitative variables and helps visualize how much of the variance is accounted for by each variable.

Scatter chart

Description automatically generated with medium confidence

The lighter the color of the variable, the more variance it captures. The variables Contract, InternetService, MonthlyCharges, and TotalCharges have the highest squared loading values. Therefore, they are more important in explaining the variance captured by the first and second dimensions than variables appearing around zero, such as Partner, Gender, PhoneService, and SeniorCitizen.

1. **Logistic Regression**

We have selected Logistic Regression as a nondescriptive method because our goal was to predict what customers would churn based on their characteristics. Churn is a dependent variable in this scenario. It is binary, meaning that it takes on two values (‘Yes’ or ’No’). The Logistic Regression is appropriate for the analysis because it predicts what group the dependent variable will fall into.

Building the Logistic Regression Model.

First, we partitioned the data into training and testing datasets. We will need the testing dataset later to evaluate the model. After partitioning, the training dataset contains 4,931 observations, and the testing dataset contains 2,112 observations.

The output of running a regression on training data is as follows:

Call:

glm(formula = Churn ~ ., family = binomial(link = “logit”), data = training)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.9235 -0.6678 -0.2817 0.7283 3.4420

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.158e+00 9.716e-01 1.192 0.23314

GenderMale 1.672e-02 7.773e-02 0.215 0.82973

SeniorCitizenYes 1.588e-01 1.016e-01 1.563 0.11801

PartnerYes -3.692e-02 9.305e-02 -0.397 0.69151

DependentsYes -1.366e-01 1.075e-01 -1.271 0.20386

**Tenure** -6.881e-02 7.951e-03 -8.654 < 2e-16 \*\*\*

PhoneServiceYes 4.705e-02 7.677e-01 0.061 0.95113

MultipleLinesYes 4.681e-01 2.099e-01 2.230 0.02572 \*

InternetServiceFiber optic 1.662e+00 9.464e-01 1.756 0.07903 .

InternetServiceNo -1.766e+00 9.595e-01 -1.841 0.06564 .

OnlineSecurityYes -2.551e-01 2.120e-01 -1.203 0.22891

OnlineBackupYes 3.872e-02 2.079e-01 0.186 0.85228

DeviceProtectionYes 2.178e-01 2.095e-01 1.040 0.29853

TechSupportYes -2.106e-01 2.117e-01 -0.995 0.31987

StreamingTVYes 6.357e-01 3.848e-01 1.652 0.09854 .

StreamingMoviesYes 5.422e-01 3.871e-01 1.401 0.16124

**ContractOne year -6.148e-01 1.303e-01 -4.720 2.36e-06 \*\*\***

**ContractTwo year** -1.278e+00 2.084e-01 -6.135 8.53e-10 \*\*\*

PaperlessBillingYes 2.365e-01 8.958e-02 2.640 0.00828 \*\*

PaymentMethodCredit card (automatic) -1.441e-02 1.368e-01 -0.105 0.91607

PaymentMethodElectronic check 2.570e-01 1.136e-01 2.262 0.02371 \*

PaymentMethodMailed check -6.463e-02 1.388e-01 -0.466 0.64157

MonthlyCharges -3.611e-02 3.767e-02 -0.959 0.33776

**TotalCharges** 3.785e-04 8.882e-05 4.262 2.03e-05 \*\*\*

Signif. Codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5707.1 on 4930 degrees of freedom

Residual deviance: 4062.7 on 4907 degrees of freedom

**AIC: 4110.7**

Number of Fisher Scoring iterations: 6

According to the Logistic Regression model, the most relevant features are Tenure, Contract, and TotalCharges. Their p-values are much lower than the 0.05 threshold. Therefore, there is an association between these variables and the probability of churn.

The Logistic Regression model included all 19 predictors from the original dataset. Let’s check if decreasing the number of variables will help us get a better-fitting model. We will use the step() function that performs variable selection in a stepwise manner, adding or removing variables based on the AIC criteria.

The algorithm selected a model containing 14 predictors. However, the AIC for the two models is very close (4110.7 vs. 4102.9).

The output of running a regression using a reduced set of predictors is as follows:

Call:

glm(formula = Churn ~ SeniorCitizen + Tenure + PhoneService +

MultipleLines + InternetService + OnlineBackup + DeviceProtection +

StreamingTV + StreamingMovies + Contract + PaperlessBilling +

PaymentMethod + MonthlyCharges + TotalCharges, family = binomial(link = "logit"),

data = training)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.8994 -0.6771 -0.2766 0.7309 3.4640

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 2.064e+00 4.179e-01 4.938 7.87e-07 \*\*\*

SeniorCitizenYes 2.339e-01 1.004e-01 2.330 0.01979 \*

Tenure -5.841e-02 7.298e-03 -8.003 1.21e-15 \*\*\*

PhoneServiceYes 9.143e-01 3.088e-01 2.961 0.00306 \*\*

MultipleLinesYes 6.685e-01 1.147e-01 5.828 5.61e-09 \*\*\*

InternetServiceFiber optic 2.659e+00 3.274e-01 8.120 4.65e-16 \*\*\*

InternetServiceNo -2.653e+00 4.029e-01 -6.584 4.58e-11 \*\*\*

OnlineBackupYes 3.183e-01 1.139e-01 2.795 0.00518 \*\*

DeviceProtectionYes 4.337e-01 1.170e-01 3.707 0.00021 \*\*\*

StreamingTVYes 9.064e-01 1.631e-01 5.559 2.71e-08 \*\*\*

StreamingMoviesYes 1.011e+00 1.641e-01 6.162 7.20e-10 \*\*\*

ContractOne year -7.153e-01 1.286e-01 -5.563 2.65e-08 \*\*\*

ContractTwo year -1.582e+00 2.165e-01 -7.309 2.69e-13 \*\*\*

PaperlessBillingYes 4.009e-01 8.903e-02 4.504 6.68e-06 \*\*\*

PaymentMethodCredit card (automatic) -7.373e-02 1.378e-01 -0.535 0.59256

PaymentMethodElectronic check 3.040e-01 1.140e-01 2.666 0.00767 \*\*

PaymentMethodMailed check -9.237e-02 1.379e-01 -0.670 0.50286

MonthlyCharges -7.865e-02 1.323e-02 -5.945 2.76e-09 \*\*\*

TotalCharges 2.912e-04 8.366e-05 3.481 0.00050 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5707.1 on 4930 degrees of freedom

Residual deviance: 4064.9 on 4912 degrees of freedom

**AIC: 4102.9**

Number of Fisher Scoring iterations: 6

We applied both models to the testing dataset. The full model had an error rate of 0.1780303. The model with the reduced set of predictors had an error rate of 0.1803977.

The two models have essentially the same error rate, which confirms that decreasing the number of predictors does not significantly improve the accuracy of the model.

Lastly, we created the ROC curve plot and calculated the area under the curve (AUC). The higher the AUC, the better predictions the model makes. The AUC ranges from 0 to 1. If it equals one, the predictions are always correct.

Chart, line chart

Description automatically generated

The AUC parameters for the two models are very close (0.8536819 and 0.8547376).

As expected, the AUC was a little higher in the second case, but the two models can be considered equal.