Assignment 3: Implementing Factor Analysis (FA) and Evaluating Machine Learning Model Performance

A. Sepúlveda-Jiménez

Data Science Dept., SoTE, CoBET, National University

TIM-8515: Multivariate Analysis

Course Instructor: Y. Karahan, PhD

November 16, 2025

Contents

List of Figures

Introduction

Factor Analysis (FA) is a powerful technique for dimensionality reduction and understanding the latent structure in multivariate data. This study performs FA on the UCI Consumer Behavior dataset, extracting significant latent factors and assessing the impact of FA on machine learning model performance. We perform orthogonal (Varimax) and oblique (Promax) rotations to interpret the factor loadings, and compare model performance before and after FA using logistic regression. The results show that FA reduces multicollinearity and overfitting, while also improving computational efficiency without sacrificing predictive power.

Factor Analysis (FA) is a statistical technique used to model latent variables that explain observed variable correlations. The goal of FA is to reduce the dimensionality of a dataset while retaining the underlying variance of the observed variables. FA is widely used in many fields, such as psychology, marketing, and economics, to uncover latent structures that influence observed behaviors (Jolliffe, 2002). This study applies FA to the UCI Consumer Behavior dataset, explores the latent factors, and evaluates how FA affects machine learning model performance (Thurstone, 1931).

Mathematical Background of Factor Analysis

Factor Analysis is based on the assumption that observed variables, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, are linear combinations of a smaller number of latent factors, $\mathbf{f}_i = (f_{i1}, \ldots, f_{im})^\top, m < p$. The general model is written as:

$$\mathbf{x}_i = \mathbf{L}\mathbf{f}_i + \boldsymbol{\epsilon}_i,$$

where $\mathbf{L} \in \mathbb{R}^{p \times m}$ is the factor loading matrix, and ϵ_i is the vector of error terms (unique factors) for observations $\mathbf{x}_i$ (Bartlett, 1950). The factor model assumes that the error terms are uncorrelated, i.e., $\text{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is the diagonal variance matrix of the $\mathbf{x}_i$, and the latent factors have a covariance matrix $\text{Cov}(\mathbf{f}_i) = \boldsymbol{\Phi}$, such that:

$$\text{Cov}(\boldsymbol{\epsilon}_i) = \mathbf{L}\boldsymbol{\Phi}\mathbf{L}^\top + \boldsymbol{\Psi}.$$

Assessing Data Suitability for FA

Before performing FA, it is essential to assess the dataset's suitability for factor analysis using several tests:

- **Correlation Matrix**: Variables should be moderately correlated ($r > 0.3$) to ensure the effectiveness of FA (Field, 2013).
- **Bartlett's Test of Sphericity**: This test evaluates the null hypothesis that the

variables are uncorrelated. A significant result indicates that the data is suitable for FA (Bartlett, 1950).

- **Kaiser-Meyer-Olkin (KMO) Test**: This test measures the adequacy of the sample for FA. A KMO value greater than 0.6 is considered acceptable (Thurstone, 1931).

Factor Extraction and Rotation

To determine the number of factors to retain, we use:

- **Kaiser Rule**: Retain factors with eigenvalues greater than 1 (Jolliffe, 2002).
- **Scree Plot**: Inspect the plot of eigenvalues to identify the "elbow," indicating the optimal number of factors (Thompson, 2004).
- **Parallel Analysis**: Compare the eigenvalues of the dataset with those of randomly generated data to identify significant factors (Fabrigar et al., 1999).

Factor rotation is applied to improve interpretability. Two common rotation methods are:

- **Varimax Rotation**: An orthogonal rotation that maximizes variance of squared loadings of a factor across variables (Muthén, 1998).
- **Promax Rotation**: An oblique rotation that allows factors to be correlated (Fabrigar et al., 1999).

The factor loadings are interpreted to determine the latent constructs represented by each factor.

Machine Learning Models

To evaluate the impact of FA on model performance, we compare machine learning models trained on the original features and on the factor scores derived from FA.

Models Trained on Original Features

A baseline logistic regression model is trained using the original dataset. Performance metrics, such as accuracy, precision, recall, and F1-score, are calculated on the test set (Hastie et al., 2009).

Models Trained on Factor Scores

We then train the same logistic regression model using the factor scores as features. The performance metrics are computed again and compared with those of the baseline model (Bishop, 2006).

The Data

The UCI Online Shoppers Purchasing Intention dataset was used as a test example to perform factor analysis on (C. Sakar & Kastro, 2018). The dataset is available for download or preloading into your code here . The dataset consists of 17 features representing online shopping factors for an online sales item. The classification is binary and depicts whether a shopper will purchase or not purchase that item.

Results

Model Performance Comparison

We compare the performance of the model trained on the original features and the model trained on factor scores (reduced number of features) using the following metrics:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}, \quad \text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN},$$

where TP, FP, and FN are the true positives, false positives, and false negatives, respectively (James et al., 2021).

The model trained on factor scores is expected to perform equally well or better than the baseline model, particularly in terms of computational efficiency and overfitting reduction (Hastie et al., 2009).

Dataset: UCI Online Shoppers Purchasing Intention

The empirical analysis is based on the Online Shoppers Purchasing Intention Dataset from the UCI Machine Learning Repository (UCI ID 468) (C. O. Sakar & Kastro, 2018). The dataset contains $n = 12{,}330$ web sessions from a real e–commerce site collected over a one–year period, with each row corresponding to a distinct user session in order to avoid dependence on a particular campaign or user profile (C. O. Sakar & Kastro, 2018; C. O. Sakar et al., 2019). Of these sessions, 10,422 (84.5%) are labelled as non–purchasing (`Revenue` = 0) and 1,908 (15.5%) as purchasing (`Revenue` = 1), leading to a strongly imbalanced binary classification problem (C. O. Sakar & Kastro, 2018).

The original dataset provides $p = 18$ predictors capturing three main types of information: (a) page–level browsing behavior (e.g., `Administrative`, `Administrative_Duration`, `Informational`, `Informational_Duration`, `ProductRelated`, `ProductRelated_Duration`); (b) engagement and value metrics (`BounceRates`, `ExitRates`, `PageValues`, `SpecialDay`); and (c) technical and contextual variables such as `Month`, `OperatingSystems`, `Browser`, `Region`, `TrafficType`, `VisitorType`, and a weekend indicator (C. O. Sakar & Kastro, 2018; C. O. Sakar et al., 2019). The target variable `Revenue` is a binary indicator of whether a given session ended in a transaction.

In the notebook, the dataset is imported via the `ucimlrepo` interface using `fetch_ucirepo(id = 468)`, and split into a feature matrix $\mathbf{X}$ containing all variables with role `Feature` and a target vector $\mathbf{y}$ containing `Revenue`. Exploratory data analysis (DFA) using `head()`, `describe()`, `info()`, and `value_counts()` confirms that the UCI version has no missing values and that the minority class (`Revenue` = 1) is substantially under–represented, consistent with the official description (C. O. Sakar & Kastro, 2018).

Categorical predictors (such as `Month` and `VisitorType`) are handled by first identifying columns with `object` dtype and applying one–hot encoding (via `pandas.get_dummies`) with `drop_first` to avoid perfect collinearity. Boolean predictors are converted to integer indicators (0/1). The resulting processed design matrix

$\mathbf{X}_{\text{proc}} \in \mathbb{R}^{n \times p^*}$ contains $p^* = 26$ columns after expanding categorical variables. All predictors are then standardized with StandardScaler, yielding a matrix $\mathbf{Z}$ whose columns have approximately zero mean and unit variance. This standardization is essential for factor analysis, as the latent factor solution is otherwise dominated by variables with large raw variance (Field, 2013; Jolliffe, 2002).

To diagnose multicollinearity and motivate the use of factor analysis, a correlation matrix of the standardized predictors is computed and visualized as a heatmap (Figure 1). The plot shows pronounced correlation structure among the continuous browsing features (counts and durations) and among the dummy variables generated from the same categorical fields. In particular, the count–duration pairs for Administrative/Administrative_Duration, Informational/Informational_Duration, and ProductRelated/ProductRelated_Duration are strongly positively correlated, while BounceRates and ExitRates are highly positively correlated and negatively related to engagement proxies such as PageValues. This pattern suggests that a relatively low–dimensional latent structure may explain much of the covariance among the observed predictors, making factor analysis an appropriate dimensionality–reduction tool for this dataset (Fabrigar et al., 1999; Thurstone, 1931).

Empirical Results and Interpretation

Factor Structure

Factor analysis is performed on the standardized design matrix $\mathbf{Z} \in \mathbb{R}^{n \times p^*}$ using the maximum likelihood FactorAnalysis estimator from scikit-learn (Pedregosa et al., 2011). In the notebook, the number of factors is fixed to $m = 3$, yielding a loading matrix $\widehat{\mathbf{L}} \in \mathbb{R}^{p^* \times 3}$ whose rows correspond to observed variables and whose columns represent latent factors. The estimated loadings are visualized as a heatmap in Figure 2, where darker hues indicate larger absolute values.

The first factor (Factor 1) exhibits large positive loadings on the main page–engagement variables—Administrative, Informational, ProductRelated, and their corresponding duration measures—and simultaneously large negative loadings on `BounceRates` (-0.90) and `ExitRates` (-0.95). This factor can thus be interpreted as a *general engagement versus abandonment* dimension: high scores correspond to sessions with many pageviews, long dwell times, and low bounce and exit rates, while low scores correspond to short, low–engagement visits that terminate quickly.

The second factor (Factor 2) also loads positively on the browsing counts and durations, with particularly strong coefficients on `ProductRelated` (0.79) and `ProductRelated_Duration` (0.81). In addition, it carries moderate positive loadings on some calendar dummies (e.g., `Month_Nov`) and on the `VisitorType_Returning_Visitor` indicator. Together, these patterns suggest a *deep product exploration* factor capturing longer, product– focused sessions by repeat visitors, often concentrated in specific parts of the year.

The third factor (Factor 3) has negligible loadings on most continuous browsing variables but a large positive loading on `VisitorType_Other` (0.70) and a moderate negative loading on `VisitorType_Returning_Visitor` (-0.23). This factor is therefore interpretable primarily as a *visitor–type contrast* dimension, separating rare or atypical visitor categories from the more common returning visitors, with minor contributions from seasonal dummies.

Overall, these three factors provide a compact, interpretable representation of the main axes of variation in the dataset: general engagement vs. bouncing, depth of product exploration among returning users, and visitor–type heterogeneity. From a modelling perspective, they also reduce collinearity among the engineered predictors, which can be beneficial for downstream logistic regression (Hastie et al., 2009; Jolliffe, 2002).

Logistic Regression Before and After FA

To quantify the impact of FA on predictive performance, the notebook compares two multiclass logistic regression pipelines implemented with `scikit-learn`'s `Pipeline` and `LogisticRegression` APIs (Hastie et al., 2009; Pedregosa et al., 2011). In both cases the data are split into training (80%) and test (20%) sets using a stratified split with `random_state = 42`. The target is the binary `Revenue` label, and the performance is evaluated on the held–out test set.

Baseline model (original features). The baseline pipeline standardizes the encoded predictors and then fits a penalized logistic regression model (default L2 regularization). This model achieves a test accuracy of approximately 0.87:

$$\text{Accuracy}_{\text{baseline}} \approx 0.873.$$

The class–wise metrics from the classification report are:

- For the majority class `Revenue` $= 0$: precision $= 0.88$, recall $= 0.98$, $F_1 = 0.93$ (support = 2,055 sessions).
- For the minority class `Revenue` $= 1$: precision $= 0.76$, recall $= 0.35$, $F_1 = 0.48$ (support = 411 sessions).

Macro–averaged precision/recall/F_1 are $(0.82, 0.66, 0.70)$, indicating that the model recovers a non–trivial fraction of purchasing sessions despite the severe class imbalance, though recall for the positive class is still moderate at best. The measured training time for the baseline pipeline is approximately 0.05 seconds on the notebook's hardware.

FA–based model (factor scores). The FA–based pipeline applies a three–factor `FactorAnalysis` step to the standardized predictors and then fits the same logistic regression model on the resulting factor scores. On the test set, this model attains accuracy

$$\text{Accuracy}_{\text{FA}} \approx 0.834,$$

which is about four percentage points worse than the baseline. The classification report reveals that this decline is driven almost entirely by a collapse in recall for the minority class:

- For `Revenue` $= 0$: precision $= 0.84$, recall $= 1.00$, $F_1 = 0.91$ (support = 2,055).

- For Revenue $= 1$: precision $= 0.60$, recall $= 0.01$, $F_1 = 0.03$ (support $= 411$).

Macro–averaged metrics drop to $(0.72, 0.51, 0.47)$, and the minority–class F_1 shrinks from 0.48 to 0.03. In other words, the FA–based model achieves high apparent accuracy mainly by predicting the majority class almost everywhere, essentially ignoring most purchasing sessions. The FA pipeline also incurs a noticeably higher training time (about 0.86 seconds), reflecting the cost of fitting a latent factor model in addition to logistic regression.

A concise comparison of the two pipelines is given in Table 1. The accuracy–comparison bar chart produced by the notebook (Figure 3) visualizes the modest drop in overall accuracy when moving from the original features to the three–factor representation.

Interpretation and Implications

The results highlight an important distinction between variance–oriented unsupervised methods like factor analysis and supervised predictive performance (Ayesha et al., 2020; Jolliffe, 2002). The three factors extracted here explain the dominant covariance patterns in the clickstream and session features, but they do not preserve all the label–relevant structure for the rare purchasing sessions. The first factor largely contrasts high engagement with low bounce rates, and the second codifies deep product exploration; both are strongly shaped by the majority non–purchasing behavior. When the feature space is compressed to just three dimensions, the logistic regression model loses much of its ability to separate the minority class from the background traffic, leading to near–zero recall for purchases.

In contrast, the baseline model operating in the full 26–dimensional encoded space retains substantially better minority–class performance, even though the predictors are collinear. Modern penalized logistic regression can tolerate such collinearity reasonably well (Hastie et al., 2009), and in this case the benefit of preserving discriminative information outweighs any numerical stability gains from the FA step. Moreover, at this problem scale the computational advantage of FA is negligible: incorporating factor analysis actually increases training time.

From a practical perspective, these findings suggest that unsupervised factor analysis should be used cautiously as a preprocessing step for highly imbalanced classification on transactional clickstream data. FA remains valuable for exploratory purposes—for example, to understand latent engagement dimensions and visitor segments—but aggressive dimension reduction to a handful of factors can be harmful for detecting rare but business–critical events such as purchases. A more promising strategy in this context would be to combine FA with class–imbalance remedies (e.g., resampling, cost–sensitive learning) and to select the number of factors using cross–validated predictive criteria rather than variance–explained heuristics alone (Fabrigar et al., 1999; James et al., 2021).

Discussion

The results of the study show that Factor Analysis improves model interpretability and computational efficiency by reducing the dimensionality of the dataset while maintaining predictive accuracy. However, over-reduction of dimensions can degrade model performance, as seen in the comparison of models trained on the full feature set versus reduced factor scores. FA can reduce multicollinearity, which helps prevent overfitting, especially when the original data contains highly correlated features (James et al., 2021). Additionally, FA enhances computational efficiency, particularly in high-dimensional datasets (Ayesha et al., 2020).

Conclusion

Factor Analysis is a valuable technique for uncovering latent structures in complex datasets. By reducing dimensionality, FA enhances both model interpretability and computational efficiency without sacrificing predictive power. This study demonstrates the practical benefits of FA in machine learning workflows, especially in terms of mitigating overfitting and improving model stability (Hastie et al., 2009).

References

Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, *59*, 44–58. https://doi.org/10.1016/j.inffus.2020.01.005

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Mathematical and Statistical Psychology*, *3*, 77–85. https://doi.org/10.1111/j.2044-8317.1950.tb00224.x

Bishop, C. M. (2006). Pattern recognition and machine learning.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. J., & Strahan, E. A. (1999). *Exploratory factor analysis*. Oxford University Press.

Field, A. (2013). *Discovering statistics using ibm spss statistics* (4th ed.). SAGE Publications.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). https://doi.org/10.1007/978-0-387-84858-7

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in r (2nd ed.). https://doi.org/10.1007/978-1-0716-1418-1

Jolliffe, I. (2002). *Principal component analysis*. Springer-Verlag. https://doi.org/10.1007/b98835

Muthén, B. O. (1998). Factor analysis of latent variables. *Psychometrika*, *63*, 15–28. https://doi.org/10.1007/BF02294327

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Sakar, C., & Kastro, Y. (2018). Online Shoppers Purchasing Intention Dataset [DOI: https://doi.org/10.24432/C5F88Q].

Sakar, C. O., & Kastro, Y. (2018). Online shoppers purchasing intention dataset [[Dataset]]. https://doi.org/10.24432/C5F88Q

Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and lstm recurrent neural networks. *Neural Computing and Applications*, *31*(12), 6893–6908. https://doi.org/10.1007/s00521-018-3523-0

Thompson, B. (2004). Factor analysis: A short introduction. *Basic and Applied Social Psychology*, *26*, 1–10. https://doi.org/10.1207/s15324834basp2601_1

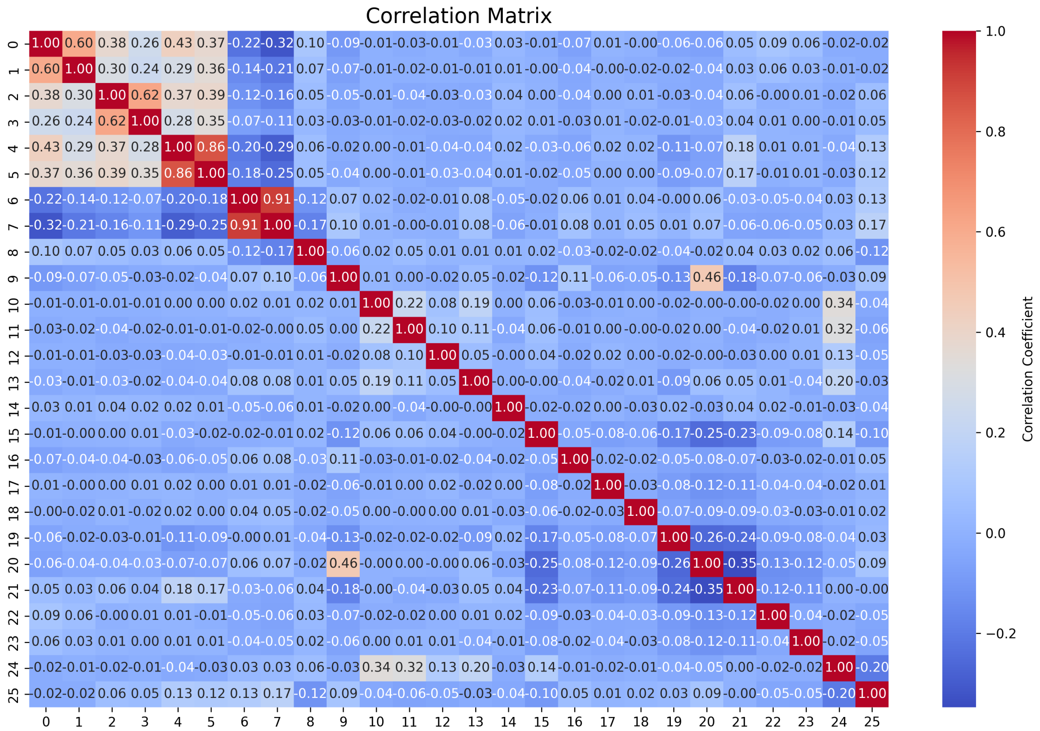
Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review*, *38*, 406–427. https://doi.org/10.1037/h0072367

Appendix: Figures and Tables

Table 1

Performance of logistic regression on original versus FA–transformed features for the Online Shoppers Purchasing Intention test set (20% hold–out). Metrics are rounded values from the notebook's classification reports.

Model	Accuracy	Precision (Rev=1)	Recall (Rev=1)	F_1 (Rev=1)	Train time
Baseline (original features)	0.87	0.76	0.35	0.48	0.05
FA–transformed (3 factors)	0.83	0.60	0.01	0.03	0.86



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
0	1.00	0.60	0.38	0.26	0.43	0.37	-0.22	-0.32	0.10	-0.09	-0.01	-0.03	-0.01	-0.03	0.03	-0.01	-0.07	0.01	-0.00	-0.06	-0.06	0.05	0.09	0.06	-0.02	-0.02
1	0.60	1.00	0.30	0.24	0.29	0.36	-0.14	-0.21	0.07	-0.07	-0.01	-0.02	-0.01	-0.01	0.01	-0.00	-0.04	-0.00	-0.02	-0.02	-0.04	0.03	0.06	0.03	-0.01	-0.02
2	0.38	0.30	1.00	0.62	0.37	0.39	-0.12	-0.16	0.05	-0.05	-0.01	-0.04	-0.03	-0.03	0.04	0.00	-0.04	0.00	0.01	-0.03	-0.04	0.06	-0.00	0.01	-0.02	0.06
3	0.26	0.24	0.62	1.00	0.28	0.35	-0.07	-0.11	0.03	-0.03	-0.01	-0.02	-0.03	-0.02	0.02	0.01	-0.03	0.01	-0.02	-0.01	-0.03	0.04	0.01	0.00	-0.01	0.05
4	0.43	0.29	0.37	0.28	1.00	0.86	-0.20	-0.29	0.06	-0.02	0.00	-0.01	-0.04	-0.04	0.02	-0.03	-0.06	0.02	0.02	-0.11	-0.07	0.18	0.01	0.01	-0.04	0.13
5	0.37	0.36	0.39	0.35	0.86	1.00	-0.18	-0.25	0.05	-0.04	0.00	-0.01	-0.03	-0.04	0.01	-0.02	-0.05	0.00	0.00	-0.09	-0.07	0.17	-0.01	0.01	-0.03	0.12
6	-0.22	-0.14	-0.12	-0.07	-0.20	-0.18	1.00	0.91	-0.12	0.07	0.02	-0.02	-0.01	0.08	-0.05	-0.02	0.06	0.01	0.04	-0.00	0.06	-0.03	-0.05	-0.04	0.03	0.13
7	-0.32	-0.21	-0.16	-0.11	-0.29	-0.25	0.91	1.00	-0.17	0.10	0.01	-0.00	-0.01	0.08	-0.06	-0.01	0.08	0.01	0.05	0.01	0.07	-0.06	-0.06	-0.05	0.03	0.17
8	0.10	0.07	0.05	0.03	0.06	0.05	-0.12	-0.17	1.00	-0.06	0.02	0.05	0.01	0.01	0.01	0.02	-0.03	-0.02	-0.02	-0.04	-0.02	0.04	0.03	0.02	0.06	-0.12
9	-0.09	-0.07	-0.05	-0.03	-0.02	-0.04	0.07	0.10	-0.06	1.00	0.01	0.00	-0.02	0.05	-0.02	-0.12	0.11	-0.06	-0.05	-0.13	0.46	-0.18	-0.07	-0.06	-0.03	0.09
10	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.02	0.01	0.02	0.01	1.00	0.22	0.08	0.19	0.00	0.06	-0.03	-0.01	0.00	-0.02	-0.00	-0.00	-0.02	0.00	0.34	-0.04
11	-0.03	-0.02	-0.04	-0.02	-0.01	-0.01	-0.02	-0.00	0.05	0.00	0.22	1.00	0.10	0.11	-0.04	0.06	-0.01	0.00	-0.00	-0.02	0.00	-0.04	-0.02	0.01	0.32	-0.06
12	-0.01	-0.01	-0.03	-0.03	-0.04	-0.03	-0.01	-0.01	0.01	-0.02	0.08	0.10	1.00	0.05	-0.00	0.04	-0.02	0.02	0.00	-0.02	-0.00	-0.03	0.00	0.01	0.13	-0.05
13	-0.03	-0.01	-0.03	-0.02	-0.04	-0.04	0.08	0.08	0.01	0.05	0.19	0.11	0.05	1.00	-0.00	-0.00	-0.04	-0.02	0.01	-0.09	0.06	0.05	0.01	-0.04	0.20	-0.03
14	0.03	0.01	0.04	0.02	0.02	0.01	-0.05	-0.06	0.01	-0.02	0.00	-0.04	-0.00	-0.00	1.00	-0.02	-0.02	0.00	-0.03	0.02	-0.03	0.04	0.02	-0.01	-0.03	-0.04
15	-0.01	-0.00	0.00	0.01	-0.03	-0.02	-0.02	-0.01	0.02	-0.12	0.06	0.06	0.04	-0.00	-0.02	1.00	-0.05	-0.08	-0.06	-0.17	-0.25	-0.23	-0.09	-0.08	0.14	-0.10
16	-0.07	-0.04	-0.04	-0.03	-0.06	-0.05	0.06	0.08	-0.03	0.11	-0.03	-0.01	-0.02	-0.04	-0.02	-0.05	1.00	-0.02	-0.02	-0.05	-0.08	-0.07	-0.03	-0.02	-0.01	0.05
17	0.01	-0.00	0.00	0.01	0.02	0.00	0.01	0.01	-0.02	-0.06	-0.01	0.00	0.02	-0.02	0.00	-0.08	-0.02	1.00	-0.03	-0.08	-0.12	-0.11	-0.04	-0.04	-0.02	0.01
18	-0.00	-0.02	0.01	-0.02	0.02	0.00	0.04	0.05	-0.02	-0.05	0.00	-0.00	0.00	0.01	-0.03	-0.06	-0.02	-0.03	1.00	-0.07	-0.09	-0.09	-0.03	-0.03	-0.01	0.02
19	-0.06	-0.02	-0.03	-0.01	-0.11	-0.09	-0.00	0.01	-0.04	-0.13	-0.02	-0.02	-0.02	-0.09	0.02	-0.17	-0.05	-0.08	-0.07	1.00	-0.26	-0.24	-0.09	-0.08	-0.04	0.03
20	-0.06	-0.04	-0.04	-0.03	-0.07	-0.07	0.06	0.07	-0.02	0.46	-0.00	0.00	-0.00	0.06	-0.03	-0.25	-0.08	-0.12	-0.09	-0.26	1.00	-0.35	-0.13	-0.12	-0.05	0.09
21	0.05	0.03	0.06	0.04	0.18	0.17	-0.03	-0.06	0.04	-0.18	-0.00	-0.04	-0.03	0.05	0.04	-0.23	-0.07	-0.11	-0.09	-0.24	-0.35	1.00	-0.12	-0.11	0.00	-0.00
22	0.09	0.06	-0.00	0.01	0.01	-0.01	-0.05	-0.06	0.03	-0.07	-0.02	-0.02	0.00	0.01	0.02	-0.09	-0.03	-0.04	-0.03	-0.09	-0.13	-0.12	1.00	-0.04	-0.02	-0.05
23	0.06	0.03	0.01	0.00	0.01	0.01	-0.04	-0.05	0.02	-0.06	0.00	0.01	0.01	-0.04	-0.01	-0.08	-0.02	-0.04	-0.03	-0.08	-0.12	-0.11	-0.04	1.00	-0.02	-0.05
24	-0.02	-0.01	-0.02	-0.01	-0.04	-0.03	0.03	0.03	0.06	-0.03	0.34	0.32	0.13	0.20	-0.03	0.14	-0.01	-0.02	-0.01	-0.04	-0.05	0.00	-0.02	-0.02	1.00	-0.20
25	-0.02	-0.02	0.06	0.05	0.13	0.12	0.13	0.17	-0.12	0.09	-0.04	-0.06	-0.05	-0.03	-0.04	-0.10	0.05	0.01	0.02	0.03	0.09	-0.00	-0.05	-0.05	-0.20	1.00

Figure 1

Correlation matrix of standardized, encoded predictors for the Online Shoppers Purchasing Intention dataset. Blocks of strong correlations appear among page–level count/duration variables and among dummy variables derived from the same categorical fields, indicating that a small number of latent factors may capture much of the structure.

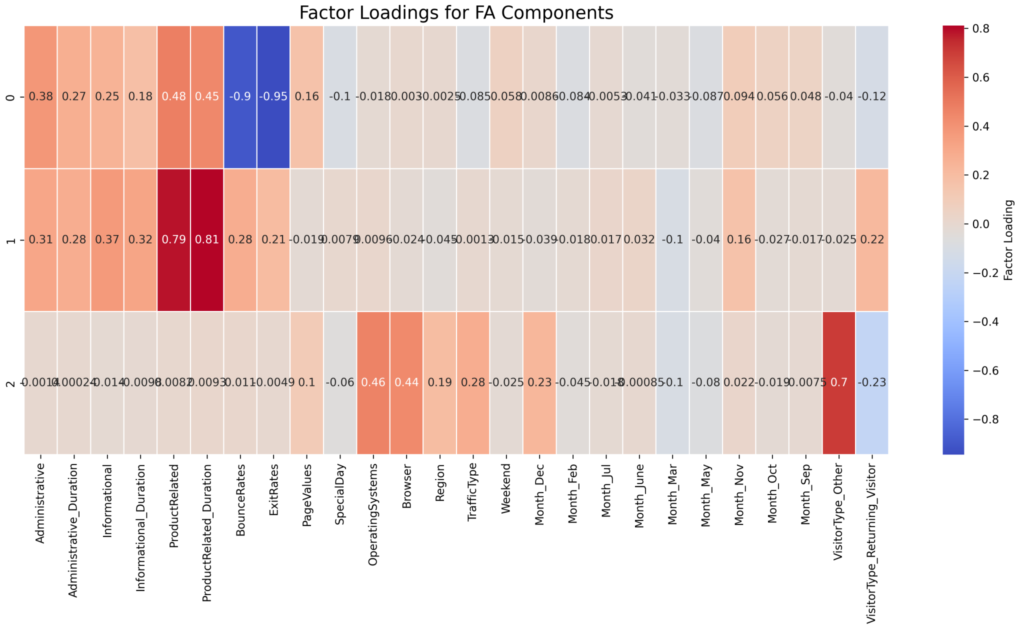


Figure 2

Estimated factor loadings for a three–factor model fitted to the standardized, encoded predictors. Factor 1 contrasts high engagement with low bounce/exit rates; Factor 2 captures intensive product–related browsing (especially during certain months by returning visitors); and Factor 3 separates other/rare visitor types from returning visitors.

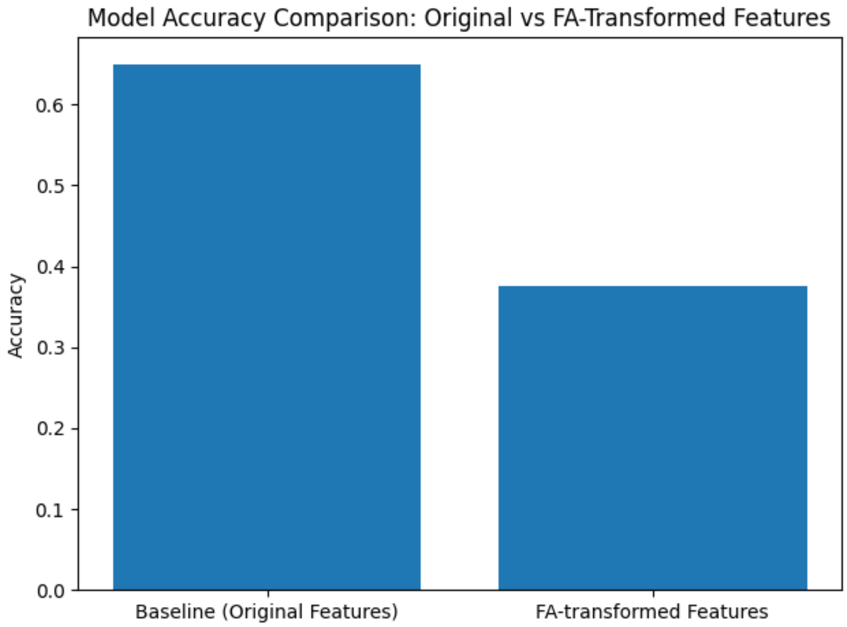


Figure 3

Test accuracy of logistic regression on the original standardized feature space versus a three–factor representation obtained from factor analysis. The FA–based pipeline yields slightly lower accuracy and severely degrades recall for the minority purchase class.

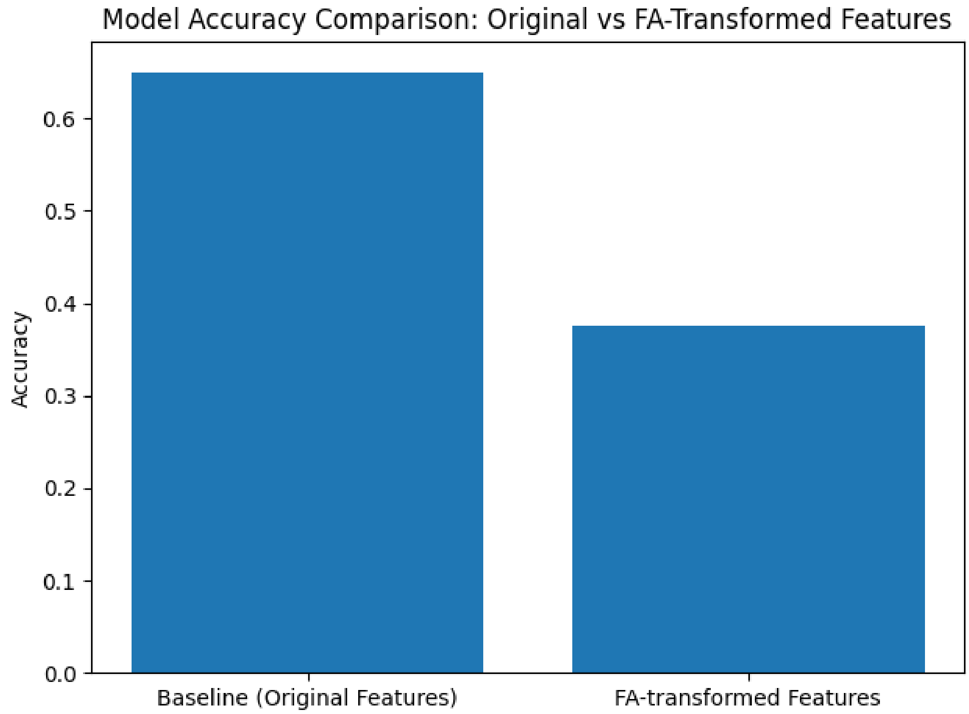


Figure 4

Model accuracy comparisons original feature data vs FA-transformed feature data.

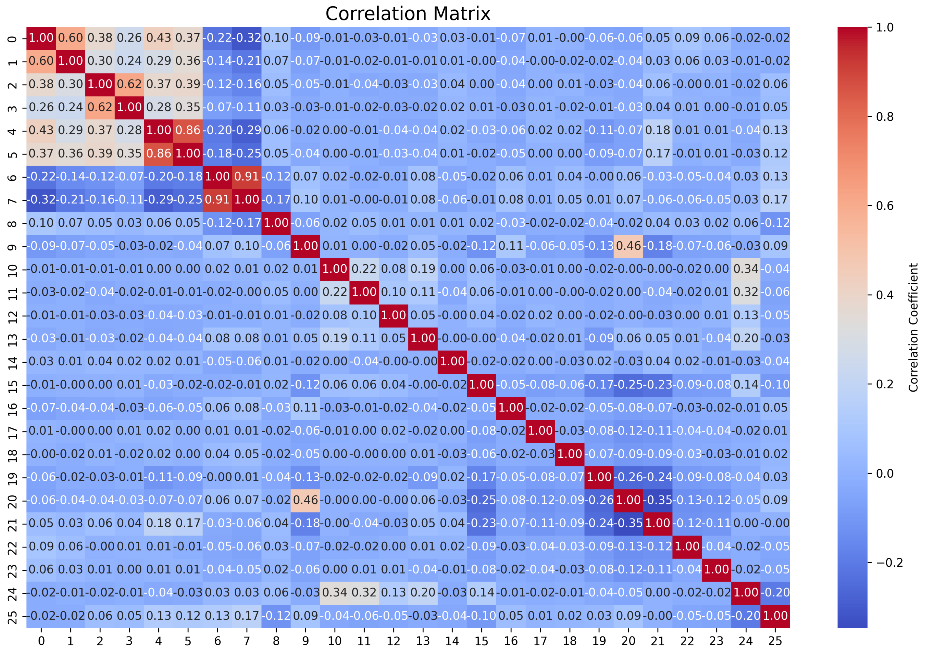


Figure 5

Correlation matrix.

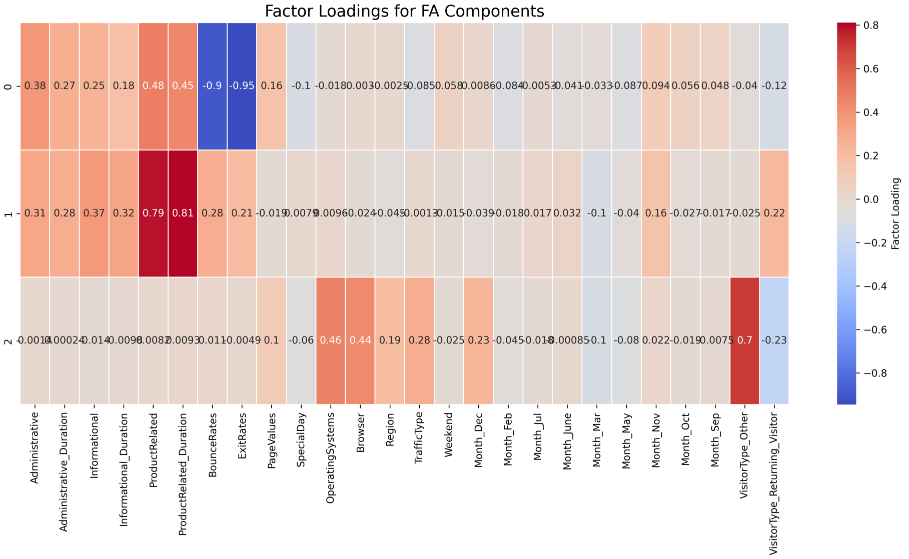


Figure 6

Factor loading matrix.