

PREDICTING RIDE-HAILING CUSTOMER LOYALTY USING RANDOM FOREST AND PRINCIPAL COMPONENT ANALYSIS

Viola¹⁾, Dewi Nasien^{2*)}, M. Hasmil Adiya³⁾

^{1,2,3} Faculty of Computer Science, Institut Bisnis dan Teknologi Pelita Indonesia, Jl. Jend. Ahmad Yani No. 78–88, Pekanbaru, Riau, Indonesia

email: ¹xooxu12@gmail.com, ²dewi.nasien@lecturer.pelitaIndonesia.ac.id,
³hasmil.adiya@lecturer.pelitaIndonesia.ac.id

*Corresponding author

Abstract

Customer loyalty is important for ride-hailing companies because users can easily change from one platform to another. This study applied Random Forest to classify the loyalty of ride-hailing users in Pekanbaru, Indonesia, and examined whether Principal Component Analysis (PCA) improved the result. The original dataset contained 200 questionnaire records covering demographic characteristics, trust, innovation, service quality, customer satisfaction, and customer loyalty. After seven duplicate records were removed, 193 records were divided into 154 training records and 39 test records. The baseline model achieved 0.67 accuracy and a weighted F1-score of 0.67. The best PCA model retained 80% of the variance, used seven trees, and achieved 0.69 accuracy with a weighted F1-score of 0.68. Service quality, customer satisfaction, trust, and innovation were the most important predictors. The model performed best for the Loyal class, while the smaller classes were harder to identify. One loyalty category was absent from the test set; therefore, the findings should be treated as an initial result rather than a model ready for operational use.

Keywords: Customer Loyalty, Ride-Hailing, Random Forest, Principal Component Analysis, Service Quality

1. INTRODUCTION

Ride-hailing services are now part of daily travel in many Indonesian cities. A user can order a vehicle, follow the driver's position, and choose a cash or digital payment method from one application. The same convenience also makes it easy to move to another provider. For that reason, a high number of transactions does not always mean that customers will continue using the same service. Earlier studies found that service quality, satisfaction, and trust are closely related to loyalty in ride-hailing services, although the strength of each relationship can differ across locations and service conditions (Pratondo & Zaid, 2021; Thaithatkul et al., 2021; Putri et al., 2024).

Research on customer loyalty often explains the relationship between variables through descriptive statistics or regression-based analysis. Those methods are useful, but a company may also need a model that can place new customers into loyalty groups. Machine-learning classification offers this type of output. Random Forest is suitable for mixed customer data because it combines several decision trees, can capture nonlinear patterns, and provides a ranking of predictor importance (Breiman, 2001). Related machine-learning studies have been used for loyalty or churn prediction in hotels, telecommunications, and subscription businesses (Hamdan & Othman, 2022; Suh, 2023; Chang et al., 2024).

A small survey dataset creates two practical problems. First, the result may change when the training and test records are divided differently. Second, a large class can dominate the evaluation and hide weak results in smaller classes. PCA can reduce correlated variables into a smaller number of components while keeping a selected amount of the original variance (Jolliffe & Cadima, 2016). However, PCA does not always improve classification because the directions with the highest variance are not necessarily the directions that best separate the classes (Wu et al., 2022). Recent reviews also identify class imbalance and limited validation as recurring weaknesses in customer-churn research (Imani et al., 2025).

This study applies Random Forest to four levels of ride-hailing customer loyalty. It has three aims: to describe the classification process, compare the model before and after PCA, and identify the variables that contributed most to the prediction. Class-level results are reported because overall accuracy alone can hide weak performance in smaller groups. This study extends previous research on service quality analysis and the application of the Random Forest algorithm by applying these approaches to the classification of customer loyalty in digital transportation services (Safarah et al., 2024; Sari et al., 2025).

2. RESEARCH METHOD

2.1 Research Design and Data

The study used a quantitative, cross-sectional questionnaire. Respondents were users of ride-hailing services in Pekanbaru, Indonesia. The original file contained 200 records. Seven duplicate records were removed, leaving 193 records for the reported modelling experiment. Four predictors described respondent characteristics, while four predictors represented customer perceptions. Customer loyalty was the target variable.

Table 1. Variables used in the classification model

Variable	Role	Measurement or categories
Gender	Predictor	Male; Female
Age group	Predictor	Respondent age category
Occupation	Predictor	Respondent occupation category
Monthly income	Predictor	Income category
Trust	Predictor	Composite score
Innovation	Predictor	Composite score
Service quality	Predictor	Composite score
Customer satisfaction	Predictor	Composite score
Customer loyalty	Target	Less Loyal; Moderately Loyal; Loyal; Very Loyal

2.2 Data Preparation

The processing steps were data inspection, duplicate checking, categorical encoding, train-test splitting, optional standardisation and PCA, Random Forest training, and model evaluation. Categorical values were converted into numeric labels. The data were divided into 80% training and 20% testing sets, producing 154 training records and 39 test records. The split used a fixed random state so the experiment could be repeated. For the PCA scenarios, standardisation was performed before component extraction. PCA was fitted on the training set and then applied to the test set to avoid information leakage.

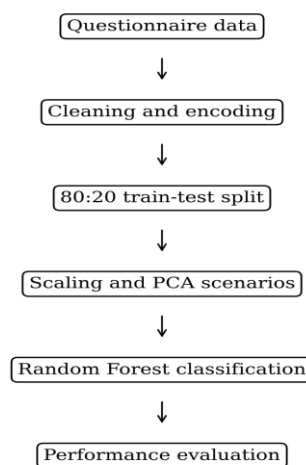


Figure 1. Data-processing and modelling workflow

2.3 Random Forest and PCA Settings

Random Forest forms an ensemble from bootstrap samples and randomly selected predictors. Each tree produces a class prediction, and the final class is selected by majority vote (Breiman, 2001). The implementation used RandomForestClassifier from scikit-learn (Pedregosa et al., 2011). In line with the source experiment, the number of trees was tested from five to eight. The best baseline setting used five trees. The best model after PCA used seven trees.

Five PCA retention levels were tested: 90%, 80%, 70%, 60%, and 50% of cumulative variance. The 80% setting retained four principal components and produced the highest test accuracy. The models were compared with accuracy, weighted precision, weighted recall, weighted F1-score, class-level precision, class-level recall, and confusion matrices. Weighted results were reported because the test classes did not contain the same number of records.

3. RESULTS AND DISCUSSION

3.1 Relationships and Predictor Importance

The correlation analysis showed that customer satisfaction had the strongest relationship with loyalty ($r = 0.83$), followed by service quality ($r = 0.79$). Gender ($r = 0.07$), age ($r = -0.06$), and monthly income ($r = -0.02$) had weak relationships with loyalty. These figures do not establish cause and effect, but they indicate that customer perceptions carried more information about loyalty than the basic demographic variables.

The Random Forest ranking produced a similar pattern. Service quality had the highest importance value (0.213), followed by customer satisfaction (0.199), trust (0.179), and innovation (0.158). Occupation, age, gender, and monthly income had lower values. Random Forest importance is model-based and should not be interpreted as a causal effect. Even so, the ranking supports the practical view that service experience is more useful for loyalty analysis than demographic grouping alone.

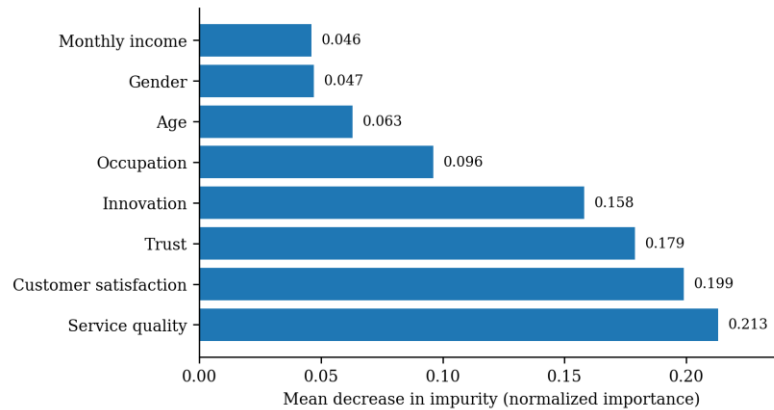


Figure 2. Feature importance in the baseline Random Forest model

3.2 PCA Retention Results

Table 2 compares the five PCA settings. Retaining 80% of the variance gave the strongest result: 0.69 accuracy, 0.69 weighted precision, 0.69 weighted recall, and a weighted F1-score of 0.68. The 90% setting reached 0.64 accuracy. Reducing the retained variance to 70%, 60%, or 50% lowered accuracy to 0.59. The result did not improve steadily as more variance was retained. This is expected because PCA preserves variance, not class separation. The 80% threshold should therefore be regarded as the best setting for this dataset, not as a general rule.

Table 2. Performance at different PCA variance-retention levels

PCA	Accuracy	Precision	Recall	F1
90%	0.64	0.69	0.64	0.65
80%	0.69	0.69	0.69	0.68
70%	0.59	0.61	0.59	0.59
60%	0.59	0.62	0.59	0.60
50%	0.59	0.62	0.59	0.60

3.3 Number of Trees

Without PCA, five trees produced the best result, with 0.67 accuracy and a weighted F1-score of 0.67. Six, seven, and eight trees did not improve the model. After PCA, seven trees gave the best result, with 0.69 accuracy and a weighted F1-score of 0.68. The improvement was only two percentage points. The changing scores across a narrow range of tree numbers also suggest that the small test set affected model stability.

Only four values of $n_{\text{estimators}}$ were examined. Further testing should include more trees and other settings, such as maximum depth, minimum samples per split, number of features, bootstrap options, and class weights. Repeated stratified cross-validation would also provide a more dependable comparison than one train-test split (Lalwani et al., 2022; Imani et al., 2025).

Table 3. Random Forest results without PCA

Trees	Accuracy	Precision	Recall	F1
5	0.67	0.68	0.67	0.67
6	0.62	0.60	0.62	0.60
7	0.59	0.59	0.59	0.59
8	0.64	0.62	0.64	0.62

Table 4. Random Forest results after PCA (80%)

Trees	Accuracy	Precision	Recall	F1
5	0.67	0.67	0.67	0.67
6	0.64	0.62	0.64	0.62
7	0.69	0.69	0.69	0.68
8	0.62	0.60	0.62	0.61

3.4 Class-Level Performance

The baseline model correctly classified 26 of the 39 test records. The Loyal class contained 26 records and was the easiest class to detect. Based on the confusion matrix, this class obtained 0.80 precision, 0.77 recall, and a 0.78 F1-score. The PCA model correctly classified 27 records and increased Loyal recall to 0.81, while precision decreased to 0.78. The smaller Moderately Loyal and Very Loyal classes remained difficult to identify.

All four loyalty categories were present in the complete dataset, but the random test set contained only Moderately Loyal, Loyal, and Very Loyal records. No Less Loyal record appeared in the test partition. Consequently, the reported accuracy does not measure the model's ability to recognize that category. The absence of one class, together with the uneven class sizes, is the main reason the model should not yet be used for automatic customer decisions. Stratified splitting, class weighting, or resampling within the training data can be tested in a larger dataset (AL-Najjar et al., 2022; Usman-Hamza et al., 2022).

Figure 3. Confusion matrices: (a) baseline model and (b) PCA model

Table 5. Class-level performance of the baseline and PCA models

Model / class	Precision	Recall	F1	n
Baseline / Moderately Loyal	0.50	0.25	0.33	4
Baseline / Loyal	0.80	0.77	0.78	26
Baseline / Very Loyal	0.42	0.56	0.48	9
PCA / Moderately Loyal	0.50	0.25	0.33	4
PCA / Loyal	0.78	0.81	0.79	26
PCA / Very Loyal	0.50	0.56	0.53	9

3.5 Discussion and Practical Use

The feature ranking gives a clearer practical message than the moderate accuracy score. Service quality should be examined first. In a ride-hailing service, this includes driver availability, waiting time, safety, fare clarity, application stability, driver response, and complaint handling. Customer satisfaction should be measured across the full service experience rather than after a single trip. Trust also matters because the application stores location, identity, travel, and payment information. Innovation is useful when a new feature makes the service easier, safer, or more reliable.

At its present stage, the model is better suited to exploratory analysis than to automatic targeting. Its output can help managers identify groups that need further study, but the prediction should be combined with transaction frequency, complaint records, promotion response, and other consent-based behavioural data. The study is limited by the small sample, one-city setting, unequal classes, one hold-out split, and a narrow parameter search. Future work should use more respondents from several cities, repeated stratified validation, stronger imbalance handling, and comparison with gradient boosting, support vector machines, and logistic regression.

4. CONCLUSION

Random Forest was applied to questionnaire data from ride-hailing users in Pekanbaru. After duplicate removal, 193 records were used in the reported experiment. The baseline model achieved 0.67 accuracy and a weighted F1-score of 0.67. With PCA retaining 80% of the variance and seven trees, accuracy increased to 0.69 and the weighted F1-score reached 0.68. Service quality, customer satisfaction, trust, and innovation were the strongest predictors, while demographic variables contributed less.

PCA produced only a small improvement and did not solve the uneven class performance. The model performed best for the Loyal class, and one loyalty category was absent from the test set. The result is therefore an initial classification study, not evidence that the model is ready for operational use. More balanced data, wider parameter testing, repeated validation, and testing on new observations are required. For service providers, the most consistent finding is the importance of reliable service, customer satisfaction, and user trust.

5. ACKNOWLEDGMENT

The authors thank the respondents who participated in the questionnaire and the Faculty of Computer Science, Institut Bisnis dan Teknologi Pelita Indonesia, for supporting the completion of the study.

6. REFERENCES

- AL-Najjar, D., Al-Rousan, N., & AL-Najjar, H. (2022). Machine learning to develop credit card customer churn prediction. *Journal of Theoretical and Applied Electronic Commerce Research*, 17(4), 1529–1542. <https://doi.org/10.3390/jtaer17040077>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chang, V., Hall, K., Xu, Q. A., Amao, F. O., Ganatra, M. A., & Benson, V. (2024). Prediction of customer churn behavior in the telecommunication industry using machine learning models. *Algorithms*, 17(6), 231. <https://doi.org/10.3390/a17060231>
- Hamdan, I. Z. P., & Othman, M. (2022). Predicting customer loyalty using machine learning for hotel industry. *Journal of Soft Computing and Data Mining*, 3(2), 31–42. <https://doi.org/10.30880/jscdm.2022.03.02.004>
- Imani, M., Joudaki, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning. *Machine Learning and Knowledge Extraction*, 7(3), 105. <https://doi.org/10.3390/make7030105>
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374, 20150202. <https://doi.org/10.1098/rsta.2015.0202>
- Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: A machine learning approach. *Computing*, 104(2), 271–294. <https://doi.org/10.1007/s00607-021-00908-y>
- 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, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pratondo, K., & Zaid, Z. (2021). Customer loyalty during pandemic: Understanding loyalty through the lens of online ride hailing service quality. *International Journal of Social Science and Business*, 5(1), 69–75. <https://doi.org/10.23887/ijssb.v5i1.30933>
- Putri, D. A., Farida, N., & Listyorini, S. (2024). Pengaruh e-service quality dan e-trust terhadap loyalitas pelanggan melalui e-satisfaction sebagai variabel intervening: Studi pada pengguna aplikasi transportasi online. *Jurnal Ilmu Administrasi Bisnis*, 13(4), 971–983. <https://doi.org/10.14710/jiab.2024.43738>
- Safarah, K. A., Inan, D. I., Juita, R., & Sirait, V. A. L. (2024). Analysing service quality using sentiment analysis and topic modeling: A case study of the Livin Mandiri application. *JOISIE (Journal of Information Systems and Informatics Engineering)*, 8(2), 209–220. <https://doi.org/10.35145/joisie.v8i2.4517>
- Sari, A., Arifin, M., & Darmanto, E. (2025). Prediksi kebutuhan stok barang menggunakan algoritma Random Forest untuk meningkatkan efisiensi penjualan. *JOISIE (Journal of Information Systems and Informatics Engineering)*, 9(2), 339–351. <https://doi.org/10.35145/joisie.v9i2.5154>
- Suh, Y. (2023). Machine learning based customer churn prediction in home appliance rental business. *Journal of Big Data*, 10, 41. <https://doi.org/10.1186/s40537-023-00721-8>
- Thaithatkul, P., Anuchitchanchai, O., Srisurin, P., Sanghatawatana, P., & Chalermpong, S. (2021). Ride-hailing applications in Bangkok: Determining service quality, passenger satisfaction, and loyalty. *International Journal of Technology*, 12(5), 903–913. <https://doi.org/10.14716/ijtech.v12i5.5247>
- Usman-Hamza, F. E., Rashid, T. A., Hassan, B. A., & Mirjalili, S. (2022). Intelligent decision forest models for customer churn prediction. *Applied Sciences*, 12(16), 8270. <https://doi.org/10.3390/app12168270>
- Wu, Z., Jing, L., Wu, B., & Jin, L. (2022). A PCA-AdaBoost model for e-commerce customer churn prediction. *Annals of Operations Research*. <https://doi.org/10.1007/s10479-022-04526-5>