Business and Organization Studies e-Journal
ISSN 3028-1938
Vol. 1 No. 2 (2023), pp. 102-118

Predicting Customer Churn in Travel and Tour Industry Using Machine Learning Algorithm Approaches

Joemarie A. Pono
Department of Economics, Sultan Kudarat State University, EJC Montilla, Tacurong City, Sultan Kudarat, Philippines
*Corresponding email: japono00456@usep.edu.ph

ABSTRACT

Predicting customer churn in the airline sector of tour and travel poses unique challenges, necessitating advanced machine learning approaches to proactively tackle dissatisfaction, optimize service reliability, and fortify loyalty within fluctuating travel patterns and preferences. This study analyzed the application of machine learning algorithms to predict customer churn in the tour and travel industry. Leveraging data obtained from Kaggle, including factors like frequent flights, annual income, and social media engagement, the study employs various classifiers and attribute selection techniques to identify key predictors of churn. Through rigorous evaluation using five-fold cross-validation, the J48 decision tree classifier emerges as the most reliable model, achieving an accuracy of 84.53% and demonstrating good agreement. The findings underscore the potential of machine learning in enabling proactive customer retention strategies and enhancing business performance in the tour and travel sector.

Business and Organization Studies e-Journal
ISSN 3028-1938
Vol. 1 No. 2 (2023), pp. 102-118

Keywords: customer churn, machine learning, SMOTE, cross-validation, classifiers

INTRODUCTION

Customer churn in the airline industry significantly impacts profitability. The ability to predict that a specific customer is at a high risk of churning represents a huge additional potential revenue source for every business. Airlines are addressing this issue by investing in advanced data analytics to understand passenger preferences, enhancing service reliability, and introducing loyalty programs to boost customer retention. Therefore, applied machine learning is an essential tool that may help to predict customer propensities to improve the overall travel experience and encourage customer loyalty, airlines that aim to reduce churn and ensure sustained business growth.

Churn prediction is the process of determining which consumers are most likely to discontinue using a business's goods or services within a given time frame. Businesses can anticipate customer attrition and take proactive steps to keep at-risk clients by examining trends and behaviors in customer data (Hadden, Tiwari, Roy, Ruta, & Research, 2007). Statistical models, machine learning algorithms, and data analytics are commonly employed in this process to identify critical churn indicators, such as decreased engagement, altered purchasing patterns, and demographic variables (Shaaban, Helmy, Khedr, Nasr, & Applications, 2012; Verbeke, Martens, Mues, & Baesens, 2011). Understanding why customers leave enables airlines to address underlying issues and improve their overall

service offerings (Mohammad, Ismail, Kama, Yusop, & Azmi, 2019).

Predicting customer churn in airline companies involves analyzing various factors, including the frequency of flights taken by passengers. Studies have shown that flight frequency can be a significant predictor of customer churn, as it reflects the engagement and loyalty of passengers. Frequent flyers are less likely to churn due to accumulated benefits and a stronger relationship with the airline. Conversely, a decline in flight frequency can signal potential churn, prompting the need for targeted retention strategies (Lee, Kim, Lee, & Systems, 2017). More so, lower-income customers are more price-sensitive. High engagement, indicated by frequent service use, generally correlates with lower churn risk (Coussement, Benoit, & Van den Poel, 2010).

In a variety of settings, including the hotel business, social media usage enhancement has changed customer behavior analysis and customer relationship management (Ahani, Rahim, & Nilashi, 2017; Garrido-Moreno, García-Morales, Lockett, & King, 2018). Through the easy exchange of customer feedback and evaluations of travel venues, destinations, and experiences, social commerce, smart tourism, and the emergence of social customers within online social communities have revolutionized the travel and hotel industries and empowered customers (Gretzel, Sigala, Xiang, & Koo, 2015; Huang, Goo, Nam, Yoo, & Management, 2017). Consequently, social media and electronic word-of-mouth have a significant impact on the performance of tourism businesses; hotels may be the most affected link in the tourism supply chain (Cantallops & Salvi, 2014; Phillips, Barnes, Zigan, & Schegg, 2017; Xie, Zhang, & Zhang, 2014)

Using machine learning and deep learning techniques, we can uncover hidden patterns and valuable insights from

complex sets of data (Krishnan, Robinson, & Chilamkurti, 2020). This approach is widely used across various fields, including predicting what customers might buy in the travel industry, foreseeing when customers might stop using services like phones or banking, and making sales and marketing strategies more effective (Chen, Wang, Zhang, & Wang, 2021). Machine learning has also been used in predicting car accidents (Mohanta et al., 2022), forecasting how much people will want to buy in farming (Chuluunsaikhan, Ryu, Yoo, Rah, & Nasridinov, 2020), detecting falls in healthcare settings (Thakur, Han, & networks, 2021), predicting when tools need to be replaced in factories, and even fixing how people sit when playing the piano (Al-Mashraie, Chung, Jeon, & Engineering, 2020; Shukla, 2021). Some studies have also looked at how machine learning can help understand why people might stop using airline services.

In churn prediction, diverse machine learning models play crucial roles. J48 Decision Trees offer interpretability by splitting data based on attributes, aiding in identifying patterns leading to churn (Wang, Xu, & Hussain, 2019). Random Forests, an ensemble method, enhances prediction accuracy by combining multiple decision trees, thereby effectively handling complex datasets (Zhang, Zeng, Zhao, Jin, & Li, 2022). Lbk, a Local Bayesian Classifier, utilizes Bayesian inference to model class probabilities, making it suitable for predicting churn by capturing the uncertainty in data (Zuin et al., 2022). Naive Bayes classifiers assume feature independence, providing fast and reliable churn predictions, especially for categorical data (Khan et al., 2024). Logistic Regression models the probability of churn using a logistic function, offering interpretable coefficients for predictor variables, thus aiding in understanding the impact of features on churn likelihood (Almeida, Etherton-Beer, Sanfilippo, & Page, 2024). Multilayer

Perceptron neural networks, with their ability to capture complex patterns and nonlinear relationships, excel in churn prediction tasks across various domains (Liu, Hu, Chen, & Control, 2023). These models, as evidenced by recent research, offer diverse yet effective approaches to churn prediction, catering to different data characteristics and business needs.

The study focused on the tour and travel customer churn prediction with the application of machine learning algorithms. While machine learning techniques are applied in various sectors for churn prediction, their application in the tour and travel domain might be relatively underexplored. Thus, there may be a lack of comprehensive studies that evaluate and develop machine learning models customized to the unique characteristics and challenges of the tour and travel industry. The predictive insights enable companies to proactively address customer dissatisfaction, enhance service offerings, and cultivate long-term relationships, ultimately fostering customer loyalty and maximizing revenue. Addressing this gap could lead to the development of accurate and industry-specific churn prediction models, thereby enhancing customer retention strategies and overall business performance in the tour and travel sector. Therefore, the study aimed to develop and implement machine learning algorithms to be able to accurately predict customer churn in the tour and travel industry.

METHOD

Dataset. The dataset of the tour and travel customer churn was obtained from Kaggle. On the data science competition platform Kaggle, users compete to develop the best models for addressing particular issues or scrutinizing

particular datasets (Puurula, Read, & Bifet, 2014). The 5 identified explanatory variables were frequent flights, annual income of travelers, synchronized social media account of travelers, and customer decision to book in a lodge or hotel using company services, and a class variable which is customer churn, otherwise (https://www.kaggle.com/datasets/tejashvi14/tour-travels-customer-churn-prediction). These attributes of the tour and travel customer churn prediction are described in Table 1.

Table 1. List of Attributes of the Tour and Travels Customer Churn Prediction

Attribute	Type	Description
FrequentFlyer	Nominal	Whether the customer takes frequent flights
AnnualIncomeClass	Nominal	Class of annual income of the user
ServicesOpted	Numeric	Number of times services opted for during recent years
AccountSyncedToSocialMedia	Nominal	Whether the company account of the user synchronized to their social media (Y), otherwise (N)
BookedHotelOrNot	Nominal	Whether the customer books lodgings/Hotels using company

		services (Y), otherwise (N)
Target	Nominal	Customer churns (1), otherwise, the customer doesn't churn (0)

Data Preparation. Data preparation and cleaning for a prediction dataset involves several critical steps to ensure the data is ready for accurate model training and evaluation. To ensure that algorithms sensitive to data scale perform optimally. the normalization of numeric attributes into a standard range, typically between 0 and 1 (*Target*). It is achieved using the *Normalize filter*. Setting the attribute class for "*Target*" is another vital step where the attribute to be predicted is specified as the class attribute in the "Classify" tab. Converting numeric attributes to nominal (*NumericToNominal*) may be necessary for certain algorithms or specific analysis needs, and this can be done using the *NumericToNominal*filter. These steps collectively enhance the quality and suitability of the tour and travel churn predictors dataset for building robust predictive models in Weka.

In the dataset of class *Target*, there are 730 instances belonging to 1, and 224 instances belonging to 0, indicating that there are imbalances in the data. The imbalanced class distribution in a training set means that some classes have significantly more instances than others. Classifier performance may suffer due to this imbalance (*Target*), as the classifiers may start to favor the majority class. It creates synthetic samples by interpolating between the chosen instance and its neighbor and makes synthetic instances for the minority class using *SMOTE* (*Synthetic Minority Over-sampling Technique*). This involves taking a weighted average of their features (Alex,

Jhanjhi, Humayun, Ibrahim, & Abulfaraj, 2022). This process helps balance the dataset, improving model performance and reducing bias towards the majority class (Bhagat & Patil, 2015).

Using *SMOTE*, the dataset was expanded by adding new synthetic samples. Undersampling the majority class (1) and oversampling in the minority class (0), the majority class reduced the number of instances in the majority class to balance the class distribution by 38.76%, while the minority class duplicated instances from the majority class to balance the class distribution for 99.55%. Consequently, the dataset has a synthetic sample of 894, reduced by 6.29%.

Selection of Attributes. The process of attribute selection employed three prominent feature selection algorithms in Weka, covering correlation-based (*CorrelationAttributeEval*), information gain-based (*InfoGainAttributeEval)*, and learning-based (WrapperSubsetEval) techniques. The WrapperSubsetEval approach identified the best number of folds for accuracy estimation, recommending a five-fold cross-validation based on the outcomes, facilitating the enhancement of attribute selection. Following the implementation of the CorrelationAttributeEval feature selection, *FrequentFlyer*(r=0.375), *AnnualIncomeClass*(r=0.233), and *BookedHotelOrNot*(r=0.206) obtained the highest correlation class, while *AccountSyncedToSocialMedia* (r=0.074) and *ServicesOpted* (r=0.056) garnered a low correlation class. Further on, using InfoGainEval, *AnnualIncomeClass*(r=0.129) and *FrequentFlyer* (r=0.127) obtained a high correlation class, while *BookedHotelOrNot*(r=0.033), *AccountSyncedToSocialMedia* (r=0.004), and *ServicesOpted*(r=0.028) obtained a low correlation class.

Data Classification and Cross-Validation. This study employs seven different classifiers to perform classification on

the training dataset. The classifiers used include *NaiveBayes, Functions.Logistic*, (a decision tree), *IBK* (k-NN), and *trees.J48* (a decision tree). Among these, are the *NaiveBayes* and *Function.Logistic* did not require any parameter tuning and were thus used with their default settings. For the *J48* classifier, confidence factors of 0.25, 0.50, and 0.75 were tested. The *IBK classifier* was tested with different values of k, specifically k-NN 3, k-NN 5, and k-NN 7. Additionally, the *WrapperSubsetEval*feature selection method was applied to determine the optimal number of cross-validation folds for accuracy estimation. The results indicated that five cross-validation folds were optimal, and thus, five-fold cross-validation was used for the training dataset (Alcala & Murcia, 2024).

To make sure the model is thoroughly tested, a five-fold cross-validation method was used for all five classifiers. This technique splits the dataset into five parts. Each time, four parts are used for training the model, and the fifth part is used to test it. This process is repeated five times, each time with a different part being tested. This helps in checking the model's accuracy better, avoids overfitting, and gives more trustworthy results on how well the model performs with different algorithms (Biol & Murcia, 2024). The validation of the five-fold cross-validation process provides a comprehensive evaluation by repeatedly training and validating the model on different subsets of the dataset, leading to more accurate predictions.

RESULTS AND DISCUSSION

The results showed that the logistic regression classifier outperforms NaiveBayes by slightly with an accuracy of 79.53% and a Kappa statistic of 0.5907, indicating moderate

agreement, while the NaiveBayes classifier correctly classifies 77.86% of instances with a moderate Kappa statistic of 0.5573, indicating a moderate agreement beyond chance. Furthermore, out of all the K-NN variations, the K-NN classifier with K=3 performs the best, having an accuracy of 84.14% and a Kappa statistic of 0.6827. The accuracy and Kappa statistics show a modest reduction with increasing K value, suggesting a slight decline in performance.

Conversely, the J48 decision tree classifier exhibits a progressive increase in accuracy and Kappa statistic when confidence factors rise. The best performance is observed at a confidence factor of 0.75, with 84.53% accuracy and a Kappa statistic of 0.6906, indicating good agreement. This implies that the J48 decision tree classifier with a confidence factor of 0.75 is the most reliable and accurate model for the given dataset among all the classifiers evaluated. The result of the classification accuracy of classifiers is displayed in Table 2.

Table 2. Classification accuracy of classifiers on the training dataset (full training)

Classifiers	Variants	Correctly classified instances (%)	K
NaiveBayes	-	742.81 (77.86%)	0.5573
Functions.Logistic	-	758.75 (79.53%)	0.5907
Lazy.IBK (K-NN)	3	802.66 (84.14%)	0.6827
Lazy.IBK (K-NN)	5	789.54 (82.76%)	0.6552
Lazy.IBK (K-NN)	7	781.36 (81.90%)	0.6381
Trees.J48	0.25	802.85 (84.16%)	0.6831
Trees.J48	0.50	805.46 (84.43%)	0.6886
Trees.J48	0.75	806.43 (84.53%)	0.6906

The results showed that while Functions.Logistics outperforms NaiveBayes but is still surpassed by K-NN and J48 versions, NaiveBayes has the lowest accuracy and Kappa statistic among the assessed models, suggesting it is less successful in this context. Furthermore, K-NN classifiers with varying values of k exhibit consistent performance, with k = 5 marginally outperforming the others. However, the model with a confidence factor of 0.75 outperforms the other J48 variations in terms of accuracy and Kappa statistic, making it the most dependable and efficient variant. Compared to NaiveBayes, Functions.Logistics, and K-NN classifiers, suggest that using the J48 variation with a confidence factor of 0.75 will probably produce the most accurate and dependable results. The result of the classification accuracy of classifiers is displayed in Table 3.

Table 3. Result of five-fold cross-validation

Classifiers	Variants	Correctly classified instances (%)	K
NaiveBayes	-	76.0769%	0.5215
Functions.Logistic	-	77.6290%	0.5526
Lazy.IBK (k-NN)	3	79.1995%	0.5840
Lazy.IBK (k-NN)	5	79.2502%	0.5850
Lazy.IBK (k-NN)	7	79.0778%	0.5816
Trees.J48	0.25	78.0657%	0.5613
Trees.J48	0.50	79.1995%	0.5840
Trees.J48	0.75	79.3212%	0.5864

.

CONCLUSION AND RECOMMENDATION

Conclusions

In classification accuracy of classifiers on the training dataset, the J48 decision tree classifier is the most reliable and accurate model for the given dataset with a confidence factor of 0.75, achieving the highest accuracy (84.53%) and a Kappa statistic indicating good agreement (0.6906). This model outperforms all other classifiers evaluated, including Naive Bayes, logistic regression, and K-NN (at various K values). It highlights the J48 decision tree as the optimal choice for achieving high classification accuracy and reliability in this context.

In addition, five-fold cross-validation is used to ensure accurate model assessment. This process is repeated five times, providing a thorough evaluation and reducing the risk of overfitting, leading to more reliable performance estimates for each classifier. The evaluation indicates that the J48 decision tree classifier with a confidence factor of 0.75 is the most effective model, achieving the highest accuracy and Kappa statistic among all evaluated classifiers. While Naive Bayes and logistic regression (Functions.Logistics) show moderate performance, the J48 classifier's superior metrics are the most reliable choice for this dataset, outperforming both K-NN variants and other models.

Recommendations

Airline companies should understand the drivers of churn that enable them to take proactive measures to enhance customer loyalty and reduce churn rates. The researcher suggests to integrate the churn prediction model with existing business intelligence and customer relationship management

(CRM) systems. Implementing the decision tree (J48) classifier significantly enhances their ability to predict and mitigate customer churn, ultimately leading to improved customer retention and business performance. It will facilitate real-time analysis and enable timely interventions to retain valuable customers. With its exceptional accuracy and reliable performance, this model offers airlines a powerful tool for tasks such as customer segmentation, demand forecasting, and personalized marketing strategies. Airlines can leverage the insights provided by this classifier to optimize their operations, improve customer satisfaction, and ultimately enhance their competitive edge in the industry, further strengthening their applicability and relevance to real-world airline operations. Therefore, the implication is that by embracing the J48 decision tree classifier, airline companies can harness the power of data-driven insights to drive success and innovation in their business strategies.

REFERENCES

Ahani, A., Rahim, N. Z. A., & Nilashi, M. (2017). Forecasting social CRM adoption in SMEs: A combined SEM-neural network method. *Computers in Human Behavior, 75*, 560-578.

Al-Mashraie, M., Chung, S. H., & Jeon, H. W. (2020). Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: A machine learning approach. *Computers & Industrial Engineering*, *144*, 106476.

Alcala, G. E. S., & Murcia, J. V. B. (2024). Machine learning techniques in employee churn prediction. *TWIST*, *19*(1), 382-387.

Alex, S. A., Jhanjhi, N. Z., Humayun, M., Ibrahim, A. O., & Abulfaraj, A. W. (2022). Deep LSTM model for diabetes prediction with class balancing by SMOTE. *Electronics*, *11*(17), 2737.

Almeida, O. P., Etherton-Beer, C., Sanfilippo, F., & Page, A. (2024). Health morbidities associated with the dispensing of lithium to males and females: Cross-sectional analysis of the 10% Pharmaceutical Benefits Scheme sample for 2022. *Journal of Affective Disorders*, *344*, 503-509.

Bhagat, R. C., & Patil, S. S. (2015, June). Enhanced SMOTE algorithm for classification of imbalanced big-data using random forest. In *2015 IEEE international advance computing conference (IACC)* (pp. 403-408). IEEE.

Biol, C. S., & Murcia, J. V. B. (2024). Supervised and Unsupervised Machine Learning Approaches in Predicting Startup Success. *TWIST*, *19*(1), 203-208.

Cantallops, A. S., & Salvi, F. (2014). New consumer behavior: A review of research on eWOM and hotels. *International Journal of Hospitality Management*, *36*, 41-51..

Chen, S. X., Wang, X. K., Zhang, H. Y., & Wang, J. Q. (2021). Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine. *Expert Systems with Applications*, *173*, 114756.

Chuluunsaikhan, T., Ryu, G. A., Yoo, K. H., Rah, H., & Nasridinov, A. (2020). Incorporating deep learning and news topic modeling for forecasting pork prices: the case of South Korea. *Agriculture*, *10*(11), 513.

Coussement, K., Benoit, D. F., & Van den Poel, D. (2010). Improved marketing decision making in a customer churn prediction context using generalized additive

models. *Expert systems with Applications*, *37*(3), 2132-2143.

Garrido-Moreno, A., García-Morales, V. J., Lockett, N., & King, S. (2018). The missing link: Creating value with social media use in hotels. *International Journal of Hospitality Management*, *75*, 94-104.

Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: foundations and developments. *Electronic Markets*, *25*, 179-188.

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, *34*(10), 2902-2917.

Huang, C. D., Goo, J., Nam, K., & Yoo, C. W. (2017). Smart tourism technologies in travel planning: The role of exploration and exploitation. *Information & Management*, *54*(6), 757-770.

Khan, M. A., Javed, K., Khan, S. A., Saba, T., Habib, U., Khan, J. A., & Abbasi, A. A. (2024). Human action recognition using fusion of multiview and deep features: an application to video surveillance. *Multimedia Tools and Applications*, *83*(5), 14885-14911.

Krishnan, C. G., Robinson, Y. H., & Chilamkurti, N. (2020). Machine learning techniques for speech recognition using the magnitude. *Journal of Multimedia Information System*, *7*(1), 33-40.

Lee, E. B., Kim, J., & Lee, S. G. (2017). Predicting customer churn in mobile industry using data mining technology. *Industrial Management & Data Systems*, *117*(1), 90-109.

Liu, X., Hu, Y., & Chen, J. (2023). Hybrid CNN-Transformer model for medical image segmentation with pyramid

convolution and multi-layer perceptron. *Biomedical Signal Processing and Control*, *86*, 105331.

Mohammad, N. I., Ismail, S. A., Kama, M. N., Yusop, O. M., & Azmi, A. (2019, August). Customer churn prediction in telecommunication industry using machine learning classifiers. In *Proceedings of the 3rd international conference on vision, image and signal processing* (pp. 1-7).

Mohanta, B. K., Jena, D., Mohapatra, N., Ramasubbareddy, S., & Rawal, B. S. (2022). Machine learning based accident prediction in secure iot enable transportation system. *Journal of Intelligent & Fuzzy Systems*, *42*(2), 713-725.

Phillips, P., Barnes, S., Zigan, K., & Schegg, R. (2017). Understanding the impact of online reviews on hotel performance: an empirical analysis. *Journal of travel research*, *56*(2), 235-249.

Puurula, A., Read, J., & Bifet, A. (2014). Kaggle LSHTC4 winning solution. *arXiv preprint arXiv:1405.0546*.

Shaaban, E., Helmy, Y., Khedr, A., & Nasr, M. (2012). A proposed churn prediction model. *International Journal of Engineering Research and Applications*, *2*(4), 693-697.

Shukla, A. (2021, July). Application of machine learning and statistics in banking customer churn prediction. In *2021 8th International Conference on Smart Computing and Communications (ICSCC)* (pp. 37-41). IEEE.

Thakur, N., & Han, C. Y. (2021). A study of fall detection in assisted living: Identifying and improving the optimal machine learning method. *Journal of sensor and actuator networks*, *10*(3), 39.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction

models with advanced rule induction techniques. *Expert systems with applications*, *38*(3), 2354-2364.

Wang, Q. F., Xu, M., & Hussain, A. (2019). Large-scale ensemble model for customer churn prediction in search ads. *Cognitive Computation*, *11*, 262-270.

Xie, K. L., Zhang, Z., & Zhang, Z. (2014). The business value of online consumer reviews and management response to hotel performance. *International Journal of Hospitality Management*, *43*, 1-12.

Zhang, G., Zeng, J., Zhao, Z., Jin, D., & Li, Y. (2022, February). A counterfactual modeling framework for Churn prediction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (pp. 1424-1432).

Zuin, G., Araujo, D., Ribeiro, V., Seiler, M. G., Prieto, W. H., Pintão, M. C., ... & Veloso, A. (2022). Prediction of SARS-CoV-2-positivity from million-scale complete blood counts using machine learning. *Communications Medicine*, *2*(1), 72.