International Journal of Scientific Research and Engineering Development-– Volume 6 Issue 1, Jan-Feb 2023

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Machine Learning for Churn Analysis and Income Classification: Techniques and Findings

Jaspreet Singh

Abstract

This manuscript presents a comprehensive analysis of machine learning techniques applied to the UCI Census Income , consisting of 48,842 records across 14 parameters. The primary objective is to determine whether an individual earns more than $50,000 annually based on their demographic and occupational information. The study explores various machine learning algorithms, employing exploratory data analysis and feature sets to evaluate their impact on model performance. Furthermore, it investigates the implications and applications of machine learning models built from the Census Income , with a focus on fairness and accuracy. The findings shed light on the advantages and limitations of leveraging machine learning algorithms for analyzing demographic and economic data in the context of financial forecasting. The manuscript concludes by identifying the most effective machine learning model across all metrics, contributing to the advancement of classification algorithms in practical settings. Researchers, practitioners, and stakeholders interested in machine learning, data analysis, and socioeconomic studies will find the UCI Census Income a valuable resource, with its wide-ranging applications and significance in multiple domains.

Introduction

Churn analysis can identify consumers who are likely to discontinue use of a product or service, thereby aiding in customer retention with minimal customer acquisition costs.It can be performed using data analysis (Kim et al., 2017), social network analysis (Phadke et al., 2013), intelligent data analysis (García et al., 2016), text analysis (Anjum et al., 2017), customer segmentation](Wu et al., 2021; Zhang et al., 2022), and traditional churn analysis (Karnstedt et al., 2010).The analysis involves preparing a churn dataset consisting of demographics, usage of services, contracts and billing, monetary value, and churn (Bach et al., 2021). The analysis can be performed using various techniques such as regression analysis, decision trees, and many other machine learning models(Bach et al., 2021; Guliyev &Tatoğlu, 2021; Wang, 2022). The results of churn analysis can help businesses to develop new products, make strategic decisions, and retain customers without loss (Guliyev &Tatoğlu, 2021).Bugajev(Bugajev, 2022)showed that the accuracy of churn prediction models is affected by the churn labeling rules used. Therefore, it is essential to have a standardized definition of churn to ensure the accuracy and comparability of churn analysis results. Another research gap is the need for more accurate and effective churn prediction models. Reichl et al (Reichl et al., 2015). mentioned that churn analysis is typically done using regression analysis, neural networks, or decision trees. However, these models may not always provide accurate results due to the complexity of churn behavior. Therefore, there is a need for more advanced and sophisticated churn

prediction models that can capture the complex nature of churn behavior. Furthermore, there is a need to consider the heterogeneity of churn customers(Park & Ahn, 2022). Park and Ahn concluded that voluntary and involuntary churn occurred due to intrinsic and extrinsic motivation. Therefore, it is essential to consider the different factors that contribute to churn behavior and develop churn prediction models that can capture these factors. Another research gap is the need for more empirical research on churn behavior in different business sectors. Lai and Zeng [4](Lai & Zeng, 2014)mentioned that their study was limited to customers from one digital library, and some hypotheses were not strictly proven due to the absence of relevant empirical research. Therefore, there is a need for more empirical research on churn behavior in different business sectors to provide more accurate and reliable churn analysis results.The objective of this study is to examine the UCI Adult dataset, perform comprehensive data analysis, develop and evaluate machine learning algorithms for income classification. The primary goal is to demonstrate the accuracy with which income can be predicted using classification algorithms, thereby aiding in the practical application of these algorithms in real-world scenarios. The study aims to provide valuable insights for researchers and practitioners in the fields of machine learning and data analysis, showcasing the significance and utility of the UCI Adult dataset across various domains.

Methodology:

This study utilizes the UCI Census Income dataset, which is available as part of the UCI Machine Learning Repository. The dataset was initially derived from a database compiled by the US Census Bureau in 1994 and has since become a widely recognized benchmark for evaluating machine learning methods.To access the dataset, researchers can visit the UCI Machine Learning Repository website, which provides a comprehensive description of the dataset's features. Additionally, the website offers various file formats, including CSV and ARFF, for download.Prior to analysis, the dataset undergoes preprocessing to ensure data quality and accuracy. This includes handling missing values, converting categorical variables into numerical representations, and scaling the features. The preprocessed dataset is then subjected to exploratory data analysis to gain insights into the distribution, relationships, and statistical properties of the variables.Furthermore, feature selection techniques, such as the mutual information criterion, are applied to identify the most relevant features for income prediction. This process aims to reduce dimensionality and improve the performance of the subsequent machine learning models.The data is split into training and testing sets using a 70-30 ratio to evaluate the performance of four different machine learning algorithms: logistic regression, decision tree, random forest, and support vector machine. The evaluation metrics employed include accuracy, precision, recall, and F1-score, providing a comprehensive assessment of each model's performance.

Throughout the methodology, reference is made to the documentation provided by the UCI Machine Learning Repository, which offers detailed information about the dataset and the preprocessing techniques employed to ensure data quality.

The link to access the UCI Census Income dataset and its associated resources can be found at: [https://archive.ics.uci.edu/ml/datasets/census+income]. Researchers are encouraged to refer to this repository for further information on the dataset and its preprocessing steps.

Key findings:

In our analysis of the UCI Census Income Dataset, we evaluated four machine learning techniques: logistic regression, random forest, and support vector machines (SVM).

The logistic regression model achieved an accuracy of 0.8451, correctly identifying around 72.84% of individuals with incomes exceeding $50,000. It demonstrated a recall of 0.5992, capturing 59.92% of all individuals with higher incomes. The F1 score of 0.6475 suggests a reasonable balance between precision and recall for this model.

The random forest algorithm yielded an accuracy of 0.8623, correctly identifying approximately 62.92% of individuals earning more than $50,000. Notably, it exhibited a higher recall of 0.7723, indicating a better ability to detect individuals with higher incomes. The corresponding F1 score of 0.6938 highlights the model's overall performance.

The SVM model achieved an accuracy of 0.8437, accurately identifying around 73.09% of individuals with incomes above the specified threshold. It successfully detected 58.51% of all individuals with higher income levels, as indicated by a recall score of 0.5851. The F1 score of 0.6499 demonstrates a harmonious balance between precision and recall for the SVM model.In summary, our analysis presents a comprehensive evaluation of these machine learning techniques on the UCI Census Income Dataset. These models exhibit varying degrees of accuracy, precision, recall, and F1 scores. Understanding the strengths and limitations of each approach provides valuable insights into factors contributing to annual income levels exceeding $50,000. This knowledge can inform policies and interventions aimed at reducing income disparities and promoting economic equity.

Significance and implications

The results of our analysis reveal that Random Forest and Logistic Regression models outperform the Decision Tree and SVM models in predicting annual income levels based on the UCI Census Income Dataset. With superior accuracy, precision, recall, and F1 scores, Random Forest stands out as the most effective model for identifying individuals with incomes exceeding $50,000. SVM and Logistic Regression models also exhibit good accuracy but may not match the recall performance of Random Forest. While the Decision Tree model provides useful insights, its lower recall suggests limitations in capturing higher income levels accurately. These findings highlight the significance of utilizing advanced machine learning techniques to understand the complex relationship between socio-economic factors and income disparities.

Conclusion

Churn analysis is valuable for customer retention and minimizing acquisition costs. Various techniques, such as data analysis, social network analysis, and customer segmentation, can be employed. However, research gaps exist in standardized churn definitions, accurate prediction models, capturing heterogeneity, and empirical research in different sectors. In the income classification study, logistic

regression and random forest models outperformed decision trees and support vector machines. Random forest demonstrated the highest accuracy, recall, and F1-score for predicting incomes exceeding $50,000. Understanding these machine learning techniques sheds light on income disparities and informs interventions for economic equity.

Anjum, A., Usman, S., Zeb, A., Afridi, I., Shah, P. M., Anwar, Z., Anjum, A., Raza, B., Malik, A. K., & Malik, S. U. R. (2017). Optimizing Coverage of Churn Prediction in Telecommunication Industry. *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2017.080523

Bach, M. P., Vugec, D. S., & Jaković, B. (2021). Churn Management in Telecommunications: Hybrid Approach Using Cluster Analysis and Decision Trees.*Journal of Risk and Financial Management*. https://doi.org/10.3390/jrfm14110544

Bugajev, A. (2022). The Impact of Churn Labelling Rules on Churn Prediction in Telecommunications.*Informatica*. https://doi.org/10.15388/22-infor484

García, D., Nebot, À., &Vellido, A. (2016). Intelligent Data Analysis Approaches to Churn as a Business Problem: A Survey.*Knowledge and Information Systems*. https://doi.org/10.1007/s10115-016-0995-z

Guliyev, H., &Tatoğlu, F. Y. (2021). Customer Churn Analysis in Banking Sector: Evidence From Explainable Machine Learning Models.*Journal of Applied Microeconometrics*. https://doi.org/10.53753/jame.1.2.03

Karnstedt, M., Hennessy, T., Chan, J., & Hayes, C. (2010). *Churn in Social Networks: A Discussion Boards Case Study*. https://doi.org/10.1109/socialcom.2010.40

Kim, S. W., Choi, D., Lee, E.-J., & Rhee, W. (2017). Churn Prediction of Mobile and Online Casual Games Using Play Log Data.*Plos One*. https://doi.org/10.1371/journal.pone.0180735

Lai, Y., & Zeng, J. (2014). Analysis of Customer Churn Behavior in Digital Libraries. *Program Electronic Library and Information Systems*. https://doi.org/10.1108/prog-08-2011-0035

Park, W., & Ahn, H. (2022). Not All Churn Customers Are the Same: Investigating the Effect of Customer Churn Heterogeneity on Customer Value in the Financial Sector.*Sustainability*. https://doi.org/10.3390/su141912328

Phadke, C., Uzunalioglu, H., Mendiratta, V. B., Kushnir, D., & Doran, D. (2013). Prediction of Subscriber Churn Using Social Network Analysis. *Bell Labs Technical Journal*. https://doi.org/10.1002/bltj.21575

Reichl, P., Egger, S., Möller, S., Kilkki, K., Fiedler, M., Hossfeld, T., Tsiaras, C., &Asrese, A. S. (2015). *Towards a Comprehensive Framework for QOE and User Behavior Modelling*.https://doi.org/10.1109/qomex.2015.7148138

Wang, G.-Y. (2022). Churn Prediction for High-Value Players in Freemium Mobile Games: Using Random Under-Sampling.*Statistika Statistics and Economy Journal*. https://doi.org/10.54694/stat.2022.18

Wu, S., Yau, W.-C., Ong, T. S., & Chong, S.-C. (2021). Integrated Churn Prediction and Customer Segmentation Framework for Telco Business.*Ieee Access*. https://doi.org/10.1109/access.2021.3073776

Zhang, T., Moro, S., & Ramos, R. F. (2022). A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation.*Future Internet*. https://doi.org/10.3390/fi14030094

