Bias in Machine Learning Models for Customer Churn Prediction within Ecommerce Ecosystem

Lena Moroz | S3063766

Artificial Intelligence Ethics and Applications

School of Computing, Engineering & Digital Technologies
Teesside University, United Kingdom

Word count: 2122

2024

1. INTRODUCTION

Recommender systems (RS) are algorithmic tools designed to assist users in discovering items they may wish to purchase or consume, serving as a critical solution to alleviate the challenges posed by information overload. Over the past decades, the prevalence of bias within recommender systems has garnered significant attention across academia, industry, and society, resulting in producing various proposals for fair recommendation methods [1,2, 3].

In the dynamic landscape of e-commerce, where customer retention is paramount, the relationship between customer churn prediction and RS holds significant importance. As per scientific articles recommendations significantly impact Amazon purchases, contributing to 30% of their transactions. [4]. While e-commerce enterprises prioritize customer retention due to its cost-effectiveness compared to acquiring new customers [5], churn prediction identifies at-risk customers, allowing businesses to retain them through proactive measures [6]. RSs complement this by personalizing recommendations to boost customer satisfaction and loyalty.

However, biases persist within RSs, impacting recommendation accuracy across various user demographics such as age and gender. Notably, female and older users often encounter inferior recommendation outcomes, underscoring the systemic biases inherent in these systems and their implications for user experiences and perceptions [7,8].

Therefore, this report, supported by our experimental work delves into the impact of bias in Machine Learning Models for Customer Churn Prediction within e-commerce ecosystems and aims to promote fairness in RSs.

2. DATA PRE-PROCESSING AND MODEL DEVELOPMENT

2.1. Dataset

This project is based on Ecommerce customer churn prediction. The dataset for this research was downloaded from Kaggle - an online platform for the community of data scientists. It is ‘Ecommerce Customer Churn Analysis and Prediction Dataset’, which belongs to a leading online E-Commerce company. The company wants to know the customers who are going to churn, so accordingly they can approach customers to offer some promos.

The dataset consists of 2 tables with a mix of categorical and numerical features, including gender of customer, and others with originally non-binary target variables indicating a range of features from customer demographics to their purchasing behaviours and engagement metrics.

The main table contains 20 columns with a total of 5630 records containing information about customers of an e-commerce company.

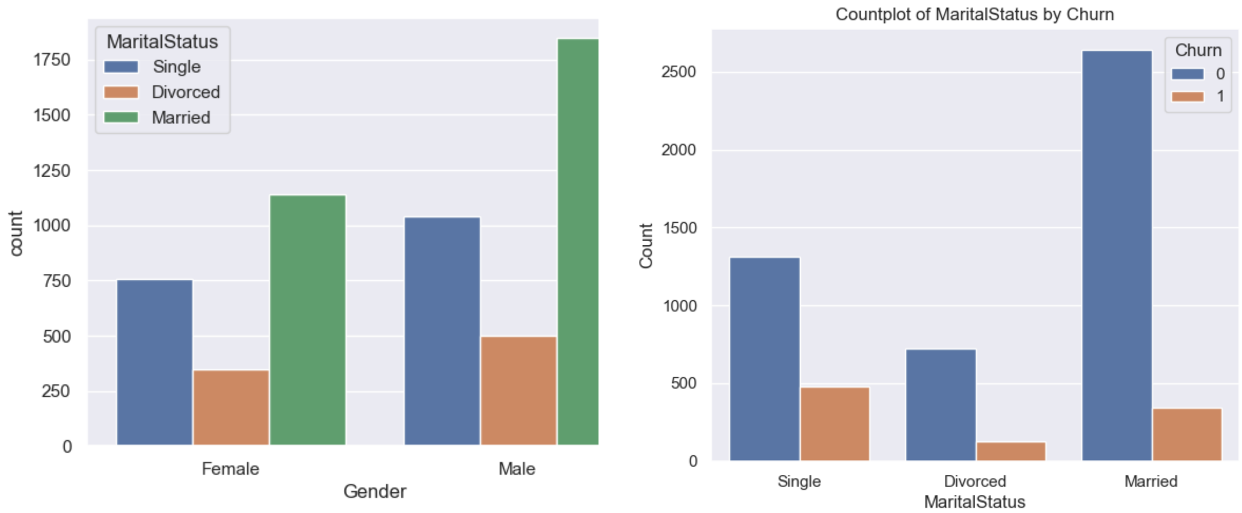
2.2. Data exploration and pre-processing stages

Firstly, we explored our data with common functions of preview, checking shape, descriptive statistics, information about data types, null values and duplicates. We dropped the column of Customer ID as it’s invaluable and even harmful for the prediction model and removed the duplicates.

After mentioned steps we got the following columns for data visualisation: 'Churn', 'Tenure', 'PreferredLoginDevice', 'CityTier', 'WarehouseToHome', 'PreferredPaymentMode', 'Gender', 'HourSpendOnApp', 'NumberOfDeviceRegistered', 'PreferedOrderCat', 'SatisfactionScore', 'MaritalStatus', 'NumberOfAddress', 'Complain','OrderAmountHikeFromlastYear', 'CouponUsed', 'OrderCount','DaySinceLastOrder', 'CashbackAmount'.

The statistical analysis below gives us brief insight into the data and makes us able to select our predictive model to conduct our research.

We visualised the marital status (protected characteristic) by gender (chosen protected characteristic for the research):

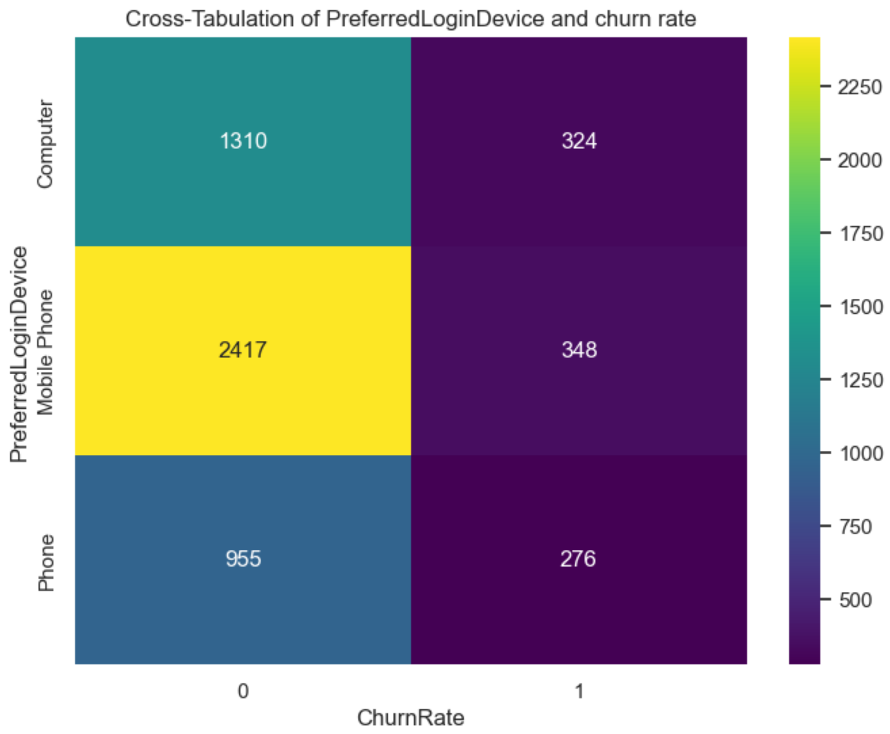


Pic. 1&2. Distribution of Marital Status by Gender and by Churn

Based on the chart above, it's evident that the least represented category is 'divorced,' whereas 'married' is the most prevalent category for both males and females, with a higher count for males.

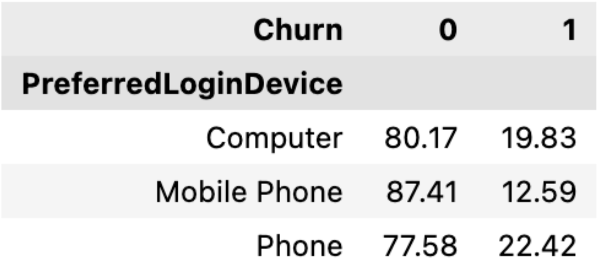
It's interesting to observe that both married and divorced individuals exhibit higher retention rates compared to singles. This suggests that individuals who have experienced marriage or divorce may have stronger ties or commitments to the products or services offered by the company. The higher retention rates among married and divorced individuals could be attributed to various factors such as shared financial responsibilities, familial obligations, or perhaps a greater sense of stability or routine in their lives.

For checking the relationship between the Preferred Login Device and Churn Rate we used chi-Square test of independence to see if there is a significant relationship:



Pic. 3. Relationship between the Preferred Login Device and Churn Rate

We calculated the percentage of churn and not churn rate for each Login Device:

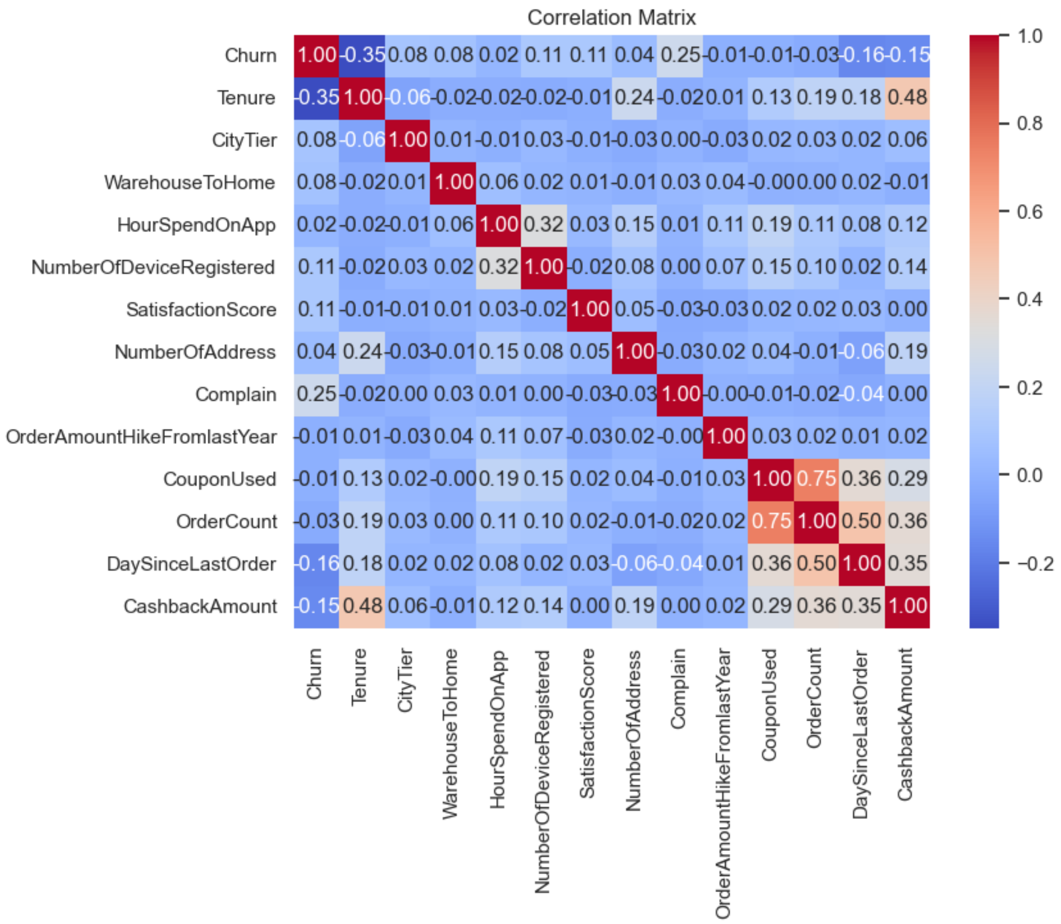


Churn	0	1
PreferredLoginDevice		
Computer	80.17	19.83
Mobile Phone	87.41	12.59
Phone	77.58	22.42

Table 1. Churn rate by Login Device

The table above shows that preferred login device has an influence on the likelihood of a customer churning, resulting in higher retention rate for mobile phones, while phones (tablets, other devices) and computers have churn rates of 22% and 20% respectively.

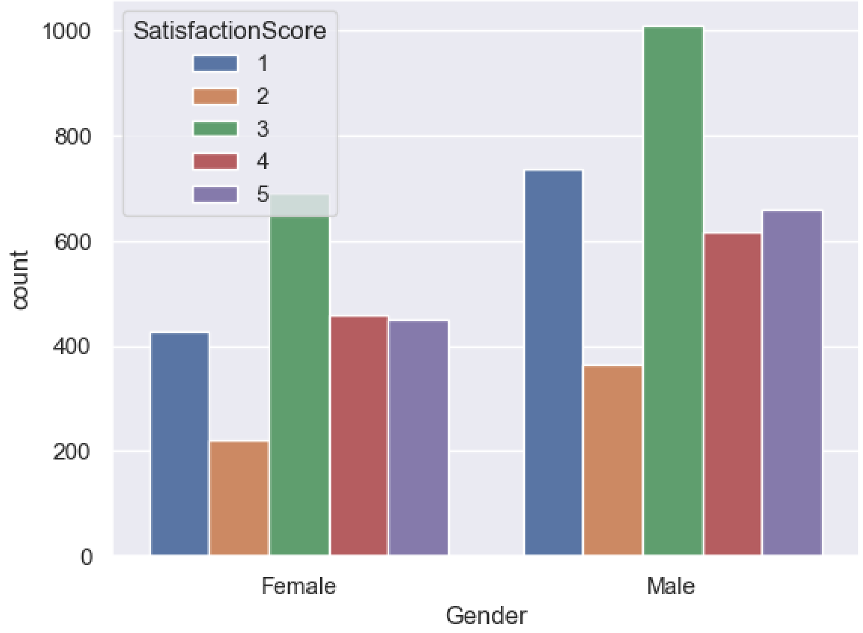
In the correlation matrix, the strongest correlation of 0.75 is observed between 'Coupon Used' and 'Order Count,' which is logically expected. The second-highest correlation, at 0.5, exists between 'Order Count' and 'Days Since Last Order.' The third-highest correlation, standing at 0.48, is found between 'Tenure' and 'Cashback Amount.'



	Churn	Tenure	CityTier	WarehouseToHome	HourSpendOnApp	NumberOfDeviceRegistered	SatisfactionScore	NumberOfAddress	Complain	OrderAmountHikeFromlastYear	CouponUsed	OrderCount	DaySinceLastOrder	CashbackAmount
Churn	1.00	-0.35	0.08	0.08	0.02	0.11	0.11	0.04	0.25	-0.01	-0.01	-0.03	-0.16	-0.15
Tenure	-0.35	1.00	-0.06	-0.02	-0.02	-0.02	-0.01	0.24	-0.02	0.01	0.13	0.19	0.18	0.48
CityTier	0.08	-0.06	1.00	0.01	-0.01	0.03	-0.01	-0.03	0.00	-0.03	0.02	0.03	0.02	0.06
WarehouseToHome	0.08	-0.02	0.01	1.00	0.06	0.02	0.01	-0.01	0.03	0.04	-0.00	0.00	0.02	-0.01
HourSpendOnApp	0.02	-0.02	-0.01	0.06	1.00	0.32	0.03	0.15	0.01	0.11	0.19	0.11	0.08	0.12
NumberOfDeviceRegistered	0.11	-0.02	0.03	0.02	0.32	1.00	-0.02	0.08	0.00	0.07	0.15	0.10	0.02	0.14
SatisfactionScore	0.11	-0.01	-0.01	0.01	0.03	-0.02	1.00	0.05	-0.03	-0.03	0.02	0.02	0.03	0.00
NumberOfAddress	0.04	0.24	-0.03	-0.01	0.15	0.08	0.05	1.00	-0.03	0.02	0.04	-0.01	-0.06	0.19
Complain	0.25	-0.02	0.00	0.03	0.01	0.00	-0.03	-0.03	1.00	-0.00	-0.01	-0.02	-0.04	0.00
OrderAmountHikeFromlastYear	-0.01	0.01	-0.03	0.04	0.11	0.07	-0.03	0.02	-0.00	1.00	0.03	0.02	0.01	0.02
CouponUsed	-0.01	0.13	0.02	-0.00	0.19	0.15	0.02	0.04	-0.01	0.03	1.00	0.75	0.36	0.29
OrderCount	-0.03	0.19	0.03	0.00	0.11	0.10	0.02	-0.01	-0.02	0.02	0.75	1.00	0.50	0.36
DaySinceLastOrder	-0.16	0.18	0.02	0.02	0.08	0.02	0.03	-0.06	-0.04	0.01	0.36	0.50	1.00	0.35
CashbackAmount	-0.15	0.48	0.06	-0.01	0.12	0.14	0.00	0.19	0.00	0.02	0.29	0.36	0.35	1.00

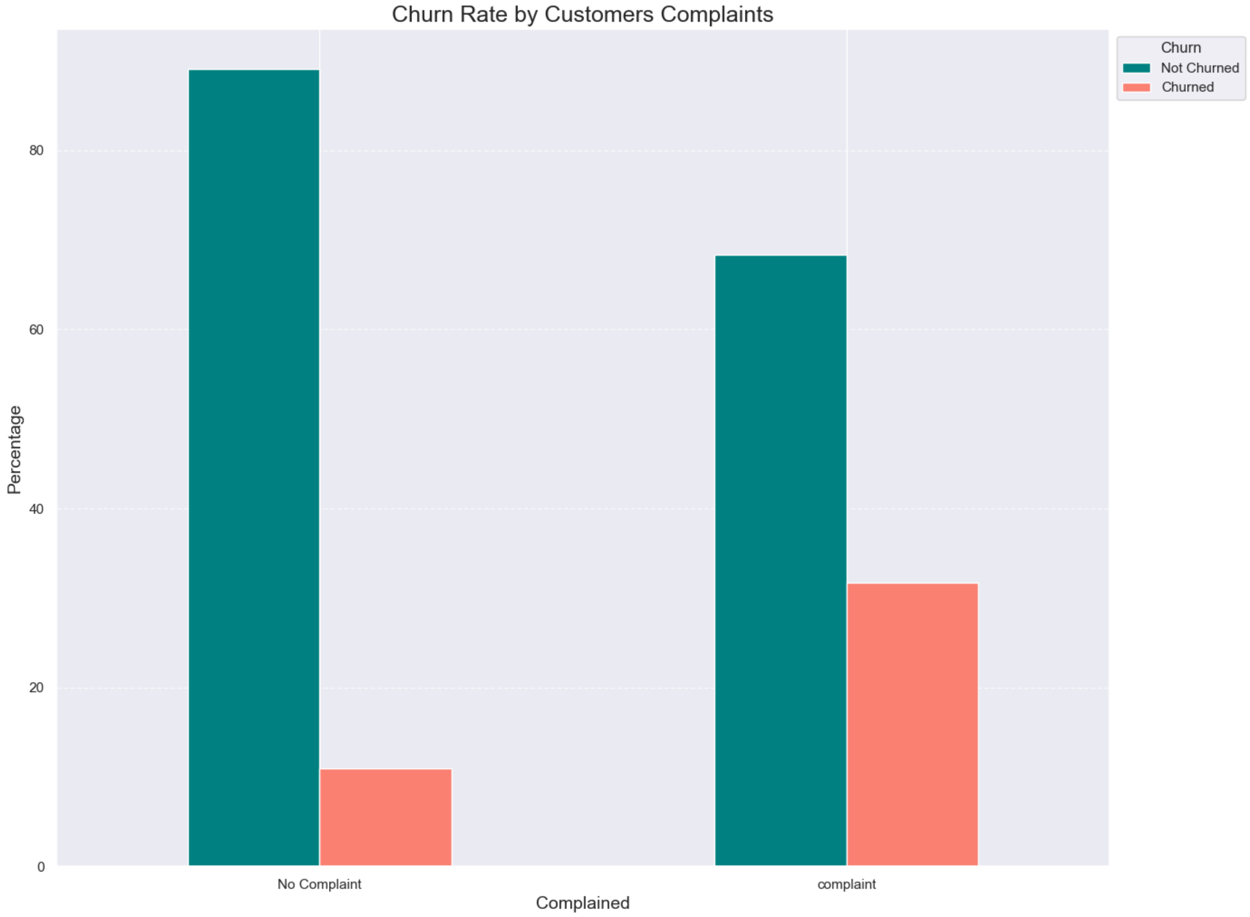
Pic. 4. Correlation Matrix between features

The distribution of ‘Satisfaction Score’ by gender shows prevalence of 3 and 1 for male, and prevalence of 3 for female, while 1,4 and 5 scores are quite balanced. The 2 score is equally less common for both male and female. If we regard 3s, 4s and 5s as ‘satisfied’ customers then more than the half of users of both genders find their experience satisfying.



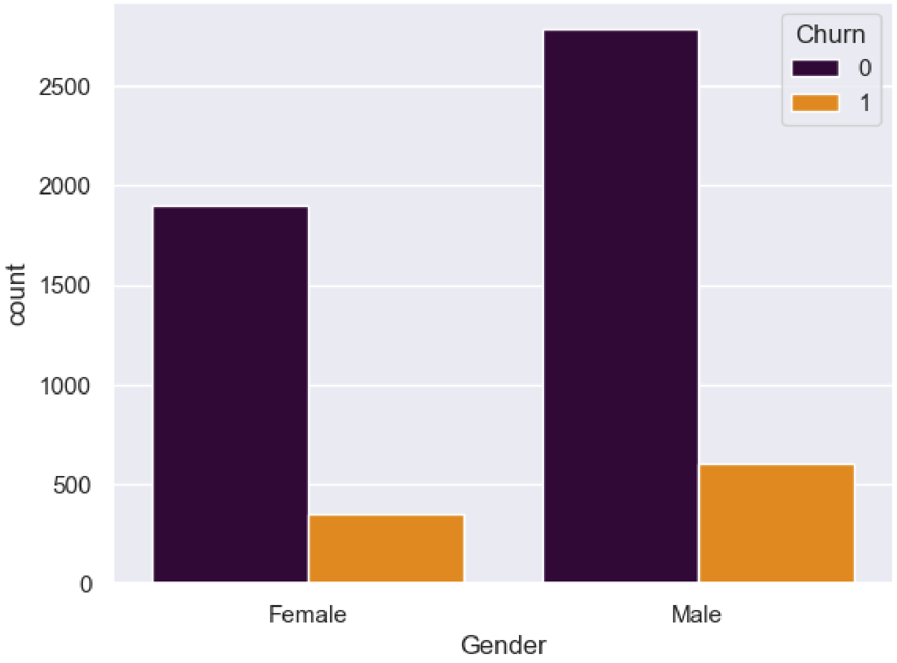
Pic. 5 Distribution of satisfaction score by gender

For analysis of the relationship between customers complain and churn rate we created a bar plot.



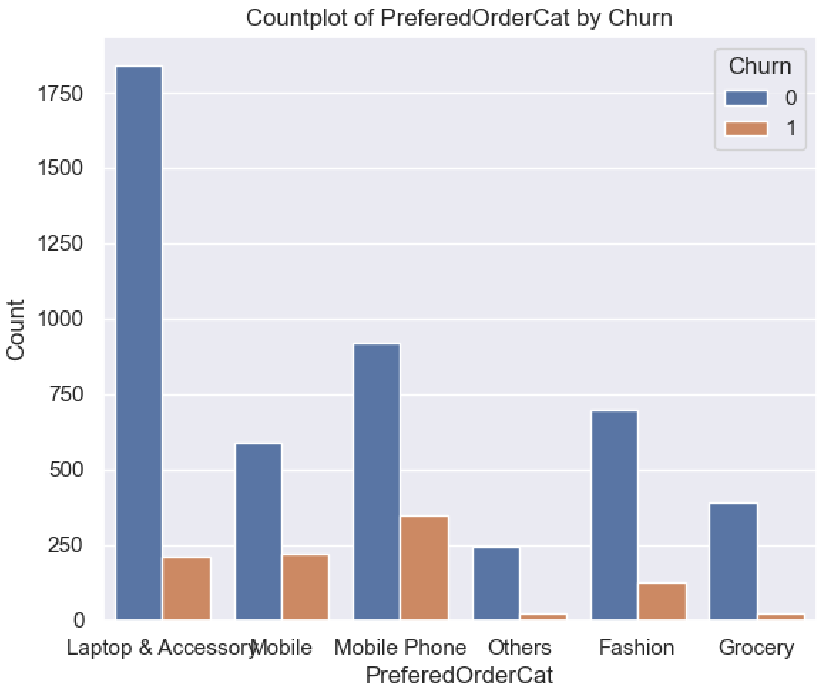
Pic. 6. Churn Rate by Customers Complains

The graph illustrates a significant data imbalance in our target variable 'Churn,' with a notably lower count of positive labels compared to negatives. Additionally, there's a slight overweight towards males. To address the issue with imbalanced data we will generate synthetic samples in next step. Regarding the slight overweight of males, it's important to assess whether this gender bias is relevant to your analysis or if it's merely reflecting the characteristics of your dataset. Thus, we chose gender label as sensitive characteristic and will use fairness-aware algorithms and metrics to ensure fairness in model predictions.



Pic. 7. Churn count by gender

The following chart reveals intriguing patterns in churn rates based on customers' preferred categories. Notably, individuals who favour categories such as 'others,' 'grocery,' and 'fashion' exhibit the lowest levels of churn. Conversely, the 'mobile phone' category stands out with an alarmingly high churn rate, nearing 45%. This substantial churn rate underscores a significant challenge within the mobile phone category, suggesting potential issues with customer retention and loyalty.



Pic. 8. Churn rate by preferred order category

Our analysis underscores the considerable impact of demographics (including gender, marital status, and city tier) and customer preferences (such as login device and number of registered devices) on churn rates. Specifically, the number of registered devices emerges as a noteworthy predictor of churn, especially when considered alongside the preferred login device. Customers with a greater number of registered devices, particularly those using computers, exhibit higher churn ratios.

2.3. Data pre-processing

For data pre-processing we handled the missing values with two different approaches. The count of missing values was between 251 and 307 instances. We divided the columns with missing values into two groups for different parts for median ('Tenure','WarehouseToHome','OrderAmountHikeFromlastYear','CouponUsed','OrderCount',' DaySinceLastOrder') and mode ('HourSpendOnApp') imputation.

For converting the categorical variables into a numerical representation, we chose One-hot encoding. After this step we divided data into features and labels. Then selected separately categorical and numerical columns.

We converted gender column into binary separately without implementation of encoding to ensure that 1 will stand for Male, and 0 for Female.

As Exploratory Data Analysis has shown the imbalance in the target variable (churn), which may result in a biased model that underperforms on the less represented class, we had to handle class imbalance. However, we decided to test ML model for both datasets, before and after handling class imbalance, for the comparison of the results.

For Addressing potential imbalance in the target variable, we chose the SMOTE (Synthetic Minority Over-sampling Technique) as it effectively enhances the representation of the minority class in the dataset, leading to more robust and accurate machine learning models [9]. It creates new instances of the minority class by interpolating between existing minority class instances, thereby increasing the representation of the minority class in the dataset. By generating synthetic samples rather than simply duplicating existing ones, SMOTE helps to avoid overfitting that can occur when the same minority class samples are repeatedly used during training.

Before training our ML model we scaled numerical data and combined it with the encoded categorical data.

2.4. ML model development process and performance evaluation metrics

We used LazyPredict library for quick evaluation of a range of machine learning models on the dataset. It automates the process of building, training, and evaluating multiple models, providing insights into how different algorithms perform on your data.

We opted for the Random Forest Classifier for our predicting customer churn due to its robust performance in handling complex datasets and ability to provide accurate predictions even with many input variables. Random Forests are well-suited for classification tasks like ours because they can effectively capture nonlinear relationships and interactions between features.

Moreover, LazyPredict shown its high accuracy of 0.98 with short training time of 0.94 seconds.

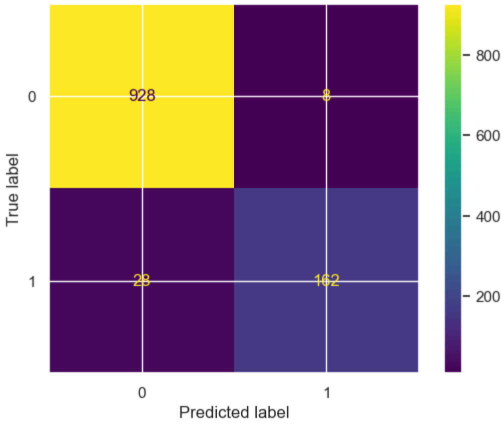
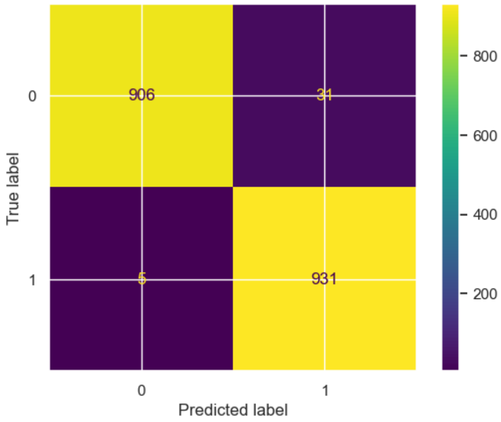
Performance metric	Without SMOTE	With SMOTE
Accuracy (% of correct classification)	0.97	0.98
Positive Rate/Precision (% of correct positive predictions)	0.95	0.97
Recall/Sensitivity/True Positive Rate (% of actual positives predicted as positive)	0.85	0.99
Correlation Matrix		

Table 2. Confusion Matrix & Performance Metrics

2.5. Gender-Aware Evaluation

To assess the effectiveness of the Random Forest Classifier for two groups distinguished by gender as the protected characteristic (males and females), we divided the data, obtained indices for "Male" and "Female" from the test dataset, and created a confusion matrix for each group.

	Without SMOTE		With SMOTE	
	Male	Female	Male	Female
Accuracy	0.96	0.98	0.98	0.98
Precision	0.95	0.96	0.96	0.98
Recall	0.83	0.89	0.99	0.99

Table 3. Comparison of gender-aware ML performance metrics

Thus, the application of SMOTE resulted in an overall improvement in the model's performance, particularly in correctly identifying positive instances (recall), while maintaining a high level of precision. This suggests that SMOTE effectively addressed class imbalance issues in the dataset, leading to more robust predictions for customer churn in e-commerce.

2.6. Fairness criteria

	Without SMOTE	With SMOTE
Equal Accuracy	0.97	0.98
Equal Positive Rate	0.95	0.97
Equality of Opportunity	0.06	0.002
Demographic Parity	0.012	0.055

Table 4. Comparison of fairness criteria for gender-aware prediction before and after SMOTE

The application of SMOTE led to improvements in equal accuracy and equal positive rate, with slight reductions in equality of opportunity and slight increase in demographic parity. Overall, the SMOTE technique contributed to a fairer and more balanced model in terms of accuracy, positive prediction rates, and fairness across different demographic groups.

The performance evaluation of the gender-aware ML model, employing the Random Forest Classifier, yields promising results:

Equal accuracy: Achieving 98% indicates a high level of balance in predictive accuracy across genders.

Equal Positive Rate: With a rate of 97%, the model demonstrates fairness in correctly identifying positive instances across gender groups.

Equality of Opportunity: A score of 0.2% indicates minimal disparity in the model's true positive rates between gender groups, signifying equitable predictive performance.

Demographic Parity: At 5.5%, the demographic parity suggests a slight imbalance in the proportion of positive outcomes between genders. While the gap exists, it remains relatively small, indicating reasonable fairness in the distribution of outcomes.

Overall, the gender-aware Random Forest Classifier exhibits commendable performance, showcasing high accuracy, balanced positive rates, and minimal disparities in predictive outcomes between gender groups. These results affirm the effectiveness of the model in providing fair and equitable predictions for customer churn in e-commerce, thereby contributing to a more inclusive and unbiased decision-making process.

3. FINDINGS

E-commerce platforms benefit significantly from proposing retention strategies for customers on the verge of leaving. Our analysis uncovers the ethical implications of the Random Forest Classifier model's performance, particularly its impact on relevant social groups, such as gender. An observation worth noting is the substantial improvement in recall when implementing the SMOTE technique, with a notable 0.14 increase in performance.

Our analysis reveals a minimal level of gender bias in the model's performance, which remains consistent even without SMOTE, as indicated by equal accuracy (97%), equal positive rate (95%), equal opportunity (6%), and demographic parity (1.2%). Post-SMOTE implementation, these metrics further improved to 98%, 97%, 0.2%, and 5.5%, respectively.

This, the initial model without SMOTE exhibited minimal bias, with relatively balanced accuracy, positive rates, and opportunity across gender groups. However, a slight imbalance in demographic parity indicated the presence of bias, albeit to a limited extent. The implementation of SMOTE led to further improvements in fairness metrics, including increased accuracy, reduced disparities in opportunity, and enhanced demographic parity. This suggests that SMOTE effectively mitigated bias in the model and improved fairness in prediction outcomes across gender groups.

Additionally, gender-aware evaluation demonstrates that SMOTE enhances performance for both genders. Looking ahead, exploring and implementing various sampling methods to address imbalanced datasets can further mitigate bias. This includes balancing not only the target value but also gender groups, thereby promoting fairness and inclusivity in model predictions.

REFERENCES

1. Yao, S. and Huang, B., 2017. Beyond parity: *Fairness objectives for collaborative filtering*. Computing Research Repository, [online] May. Available at: http://arxiv.org/abs/1705.08804 [Accessed 28 April 2024].
2. Burke, R., Sonboli, N., Mansoury, M. and Ordonez-Gauger, A., 2017. *Balanced neighborhoods for Fairness-Aware collaborative recommendation.* In FATREC Workshop on Fairness, Accountability and Transparency in Recommender Systems at RecSys. Available at: http://scholarworks. boisestate.edu/fatrec/2017/1/3/ [Accessed 28 April 2024].
3. Kamishima, T. and Akaho, S., 2017. *Considerations on recommendation independence for a Find-Good-Items task.* In Proc. of Workshop on Fairness, Accountability and Transparency in Recommender Systems at RecSys. Available at: http://scholarworks. boisestate.edu/fatrec/2017/1/11/ [Accessed 28 April 2024].
4. Linden, G., Smith, B. and York, J., 2003. *Amazon.com recommendations: item-to-item collaborative filtering*. IEEE Internet Computing, 7(1), pp.76–80. doi: 10.1109/MIC.2003.1167344.
5. Saghir, M., Bibi, Z., Bashir, S. and Khan, F.H., 2019. *Churn prediction using neural network-based individual and ensemble models*. In 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), pp. 634-639. IEEE.
6. Shobana, J.A., et al., 2023. *E-commerce customer churn prevention using machine learning-based business intelligence strategy*. Measurement: Sensors. Available at: https://www.sciencedirect.com/science/article/pii/S2665917423000648 [Accessed 28 April 2024].
7. Wang, Y., Ma, W., Zhang, M., Liu, Y. and Ma, S., 2022. *A Survey on the Fairness of Recommender Systems.* ACM Transactions on Information Systems. doi: https://doi.org/10.1145/3547333.
8. Ekstrand, M.D., Tian, M., Azpiazu, I.M., Ekstrand, J.D., Anuyah, O., McNeill, D. and Pera, M.S., 2018. *All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Proceedings of Machine Learning Research, Vol. 81), pp. 172–186. Available at: http://proceedings.mlr.press/v81/ekstrand18b.html [Accessed 28 April 2024].
9. Wu, S., Yau, W.-C., Ong, T.-S. and Chong, S.-C., 2021. *Integrated Churn Prediction and Customer Segmentation Framework for Telco Business*. IEEE Access, 9, pp.62118-62136. doi: 10.1109/ACCESS.2021.3073776.