A picture containing icon

Description automatically generated

**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Strategic Thinking |
| **Assessment Title:** | CA 2 |
| **Lecturer Name:** | James Garza ([james@cct.ie](mailto:james@cct.ie)) |
| **Student Full Name:** | [Riccardo Possieri](https://moodle.cct.ie/user/view.php?id=13536&course=2625) |
| **Student Number:** | sba23439 |
| **Assessment Due Date:** | 15th December 2023 23:59 |
| **Date of Submission:** | 26/11/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Table of contents

1. Introduction

2. Goal

3. Import Datset & libraries

4. Overview

5. EDA

6. Data Pre-processing

7. Splitting

8. Modelling

9. Oversampling

10. Conclusions

11. References

12. GitHub repo link

*Why 10% splitting*

*Speak about The box plots provided valuable insights into the distribution of key features, such as 'HourSpendOnApp' and 'SatisfactionScore,' in relation to the target variable 'Churn.'*

*Speak about the matrix correlation*

*Better English in general*

1. Introduction

The *'E Commerce Dataset'* and *'E Comm details'* Excel files have been helpful to understand and give us a comprehensive exploration into customer churn within a company. By leveraging machine learning techniques, we aimed to predict and understand the factors influencing customer attrition. This second part of capstone project revolved around a proactive approach to reduce churn and try to keep them in the company.

2. Goal

Our primary objective was to convert the *'Churn'* feature as the target variable for predictive modeling. By comparing *Churn* with other relevant features, we aspired to develop a robust machine learning algorithm capable of predicting and reducing churn for the company.

3. Import Dataset & Libraries

The initial step involves importing necessary datasets (*'E Commerce Dataset'* and *'E Comm details'*) and libraries essential for data analysis, manipulation, and machine learning. This ensures a seamless transition from data exploration to model development.

4. Overview

The *'E Commerce Dataset'* consists of 20 features and 5630 observations, with a mix of numerical and categorical columns.

Target Variable - *Churn*

No churn: 4682

Churn: 948

A blue and orange rectangular chart

Description automatically generated

Visualization

Visualizations, including box plots, were employed to discern patterns and relationships between features and the target variable. Key observations include gender-specific churn rates and city-tier-specific order counts.

*Churn* by *Preferred Payment Mode*. Debit and credit cards emerge as the most frequently utilized payment modes. However, a noticeable trend reveals that customers who opt for Cash on Delivery (COD) payment mode exhibit a significantly higher churn rate.

A graph of a bar chart

Description automatically generated with medium confidence

*Churn* by *Preferred Login Device*. The analysis of churn in relation to the preferred login device indicates that individuals who use a computer as their preferred login device demonstrate a higher percentage of churn compared to those who opt for a mobile phone as their preferred login device.

A graph of a login device

Description automatically generated

The box plots provided valuable insights into the distribution of key features, such as *'HourSpendOnApp'* and *'SatisfactionScore*,' in relation to the target variable *'Churn*.'

A comparison of a bar graph

Description automatically generated

5. EDA

In the exploration of the *'E Commerce Dataset*,' a comprehensive analysis was undertaken to understand the nuances and patterns within the data. The following conclusions have been drawn from the various aspects of the dataset.

We proceed to follow the variables description we did in the first step and analyse, handling missing values first, and pre-processing them one by one, if necessary, later. "Tenure" had 9% of zeros due to the longevity in the company and it has 4.7% of missing values which we proceeded to handle with a central tendency with the mean to maintain the overall distribution of this feature. After that, we replaced the missing values of *'WarehouseToHome'* with the median.

Consequently, we asked ourself the following questions:

Is there any realtionship between *Churn* and *Gender*?

what about *HourSpendOnApp* comparing that with *Gender*?

Which *CityTier* has the highest *OrderCount*?

An examination of the relationship between *Gender* and *Churn* revealed that since the percentage of *Churn* in male customers is slightly higher than that of female customers, we can assume that we should focus more on addressing the needs of male customers to positively affect our target variable and keep working on Female gender to increase the Total number of customers.

For the next question, the first step has been handling 4.5% of missing values with a mean and keep save the same frequency because, as the html report shows, *HourSpendOnApp* has a tendency as three hours. Then we focus on a relationship between *HourSpendOnApp* with *Gender*.

We performed the independent t-test to determine if there was a significant difference between the mean of *HourSpendOnApp* in Male & *HourSpendOnApp* in Female:

Mean *HourSpendOnApp* (Male): 2.920991368409478

Mean *HourSpendOnApp* (Female): 2.947420571972913

T-test p-value: 0.17027252509743426

There was no strong statistical evidence to suggest that the average "*HourSpendOnApp*" was significantly different between male and female customers. The p-value of 0.17027252509743426 was greater than the conventional threshold *of 0.05*, indicating that any observed differences in means could have been due to random variation. The FacetGrid showed out our analysis. This suggested us that, on average, the time spent on the mobile application was consistent across gender categories.

A graph of two people

Description automatically generated

To handle all the missing values We proceed to apply the mean to the last four variables which had missing values too and they were:

*OrderAmountHikeFromlastYear* has 265 (4.7%),

*CouponUsed* has 256 (4.5%),

*OrderCount* has 258 (4.6%),

*DaySinceLastOrder* has 307 (5.5%)

Once we didn't have any missing values and we could have focused on *CityTier* variable and comparing it with other features and answering ourselves the last question: Which *CityTier* has the highest *OrderCount*?

With a preliminary study, we saw *CityTier* 1 had the highest value with 10836,4328 orders counted.

1 10836.432800

2 627.032018

3 5471.600335

Following that, we transformed those numbers into INT values, and we got:

1 10835

2 627

3 5471

The exploration of *'CityTier'* unveiled that *CityTier 1* exhibited the highest *'OrderCount*,' indicating a higher volume of orders compared to other city tiers.

Conclusions of EDA

The exploratory data analysis not only provided insights into individual features but also laid the groundwork for subsequent modeling. The initial understanding of feature relationships and the impact of different variables on customer churn is pivotal for making informed decisions during the project's lifecycle.

As we progress into the modeling phase, the insights gained from the EDA will serve as a valuable foundation for developing a robust machine learning algorithm. By addressing gender-specific churn rates, understanding the significance of certain features, and recognizing the influence of city tiers on order counts, we are better equipped to derive meaningful conclusions from the forthcoming predictive models.

The data preprocessing steps undertaken ensure the dataset's completeness and reliability, setting the stage for accurate model training and evaluation. With these foundations in place, we transition to the modeling phase, where the predictive power of machine learning algorithms will be harnessed to forecast customer churn effectively.

6. Data Pre-processing

In the data pre-processing phase, key steps were taken to ensure the dataset's compatibility with machine learning algorithms. We converted the selected columns to integer type: *'CouponUsed'* and *'OrderAmountHikeFromlastYear'* were converted to integers to facilitate modelling. The, we did the Label encoding step. Categorical variables were label-encoded using the LabelEncoder from scikit-learn, transforming them into numerical values for machine learning compatibility.

We conducted the Correlation analysis to identify relationships between features and the target variable *'Churn*.' The correlation matrix and visualization provided valuable informations about it.

A screenshot of a computer screen

Description automatically generated

Features positively correlated with *'Churn'*: 'Complain,' *'MaritalStatus,' 'NumberOfDeviceRegistered,' 'SatisfactionScore,' 'PreferedOrderCat,' 'CityTier,' 'WarehouseToHome,' 'NumberOfAddress.*'

Features negatively correlated with 'Churn': *'Tenure,' 'DaySinceLastOrder,' 'CashbackAmount.'*

A graph with different colored bars

Description automatically generated

The heatmap visually represented these correlations, offering a comprehensive overview of feature interactions.

7. Data Splitting

The dataset was split into training and testing sets, with 10% reserved for testing. We decided to use 10% without the comparison of other percentages to not make the Assessment not to much long and keep follow the questions of the Professor. We also didn’t do the cross-validation part so we didn’t check which was the best parameter to use in each model. We chose two models: Logistic Regression and Random Forest Classifier.

8. Modelling

Logistic Regression

Logistic Regression was trained on the dataset, and its accuracy on the test set was found to be 85.26%.

Classification report metrics for Logistic Regression:

*Precision: 0.63*

*Recall: 0.28*

*F1-score: 0.39*

Random Forest Classifier

Random Forest Classifier demonstrated superior performance compared to Logistic Regression, achieving an accuracy of 97.87%.

Classification report metrics for Random Forest Classifier:

*Precision: 0.91*

*Recall: 0.71*

*F1-score: 0.80*

Model Comparison and Evaluation

Based on the provided classification reports, it looked like that the Random Forest model performed better than the Logistic Regression model, especially when considering metrics like precision, recall, and F1-score for the positive class (Churn = 1). In general, the RF model indicated better performance in identifying customers who are likely to churn.

Logistic Regression Accuracy: 85.25%

Random Forest Accuracy: 97.86%

9. Oversampling

It is an important step we decided to do only at the end because we wanted to explore and understand the natural distribution of our data. We Trained initial LR & RF models without balancing to gauge their performance and how well our model was handling the imbalanced classes and then, comparing it once we balanced our target variable. To further enhance model performance, oversampling of the minority class (Churn = 1) we performed using SMOTE (Synthetic Minority Over-sampling Technique). We resampled the dataset and standardized it before training Logistic Regression and Random Forest models.

Logistic Regression Accuracy (with oversampling): 75.13%

Random Forest Accuracy (with oversampling): 97.69%

10. Conclusions

In our exploration of model performance, we embarked on a journey to understand how class imbalance impacts the effectiveness of our predictive models. Initially, without addressing the imbalance, we trained Logistic Regression and Random Forest models on the original dataset. The results were promising, especially for the Random Forest model, which exhibited high accuracy and precision.

However, in recognizing the imbalanced distribution of our target variable, 'Churn,' we implemented the Synthetic Minority Over-sampling Technique (SMOTE) to rectify the imbalance issue. This oversampling technique aimed to provide our models with a more equitable representation of the minority class, ultimately improving their ability to identify instances of churn.

After applying SMOTE, we revisited our models to observe the impact of oversampling on their performance. The findings were intriguing.

For Logistic Regression, although the overall accuracy witnessed a slight decrease, there was a substantial improvement in recall. This indicates that the model became more adept at capturing instances of churn, a critical aspect in our context. The trade-off resulted in a balanced precision-recall dynamic.

On the other hand, the Random Forest model, even after oversampling, maintained its exceptional performance. With a high level of accuracy and a well-balanced precision-recall trade-off, it demonstrated resilience to class imbalance, further solidifying its position as the preferred choice for predicting and mitigating customer churn within the company.

In conclusion, the strategic application of oversampling techniques has proven beneficial, particularly in enhancing the recall of our models. These insights will guide us as we fine-tune our models for optimal performance and consider deployment strategies to effectively address the challenges of customer churn.

11. References

https://seaborn.pydata.org/generated/seaborn.violinplot.html

https://www.tutorialspoint.com/python\_data\_science/python\_p\_value.htm

Brownlee, J. (2020). Imbalanced Classification with Python. Machine Learning Mastery

Müller, A.C. and Guido, S. (2018) Introduction to machine learning with python: A guide for data scientists. Sebastopol: O’Reilly Media

McKinney, W. (2022) Python for Data Analysis: Data wrangling with pandas, NumPy, and Jupyter. Beijing: O’Reilly

iguazio.com (2014). What is Model Accuracy in Machine Learning. [online] Iguazio. Available at: https://www.iguazio.com/glossary/model-accuracy-in-ml/ [Accessed 2 December 2023].

12. GitHub repo link

https://github.com/riccardopossier/CA2\_Strategic\_Thinking\_Possieri\_Riccardo/tree/main