E-commerce Customer Churn Analysis

CAPSTONE PROJECT 3
PRISCILLA HILARY K
JCDS-0310 PURWADHIKA JOGJA

May 27, 2024

CONTENTS

Bussines Problem

WHAT IS THE BACKGROUND OF THE ANALYSIS?

In recent years, the E-commerce company has experienced good customer growth, but recent data shows an increase in the number of customers who **Churn** from the E-Commerce website.

To retain profitability growth, the company can employ two strategies:

1. Retain existing customers
2. Acquire new customers

In research findings that customer acquisition costs are **five times** higher than customer retention costs.

Bussines Problem

PROBLEM STATEMENT

One of the challenges faced by e-commerce businesses is to retain customers and ensure they continue making transactions.

GOALS

In many business scenarios, retaining customers or detecting customers who are likely to switch to competitors can be more important than attracting new customers. In this case, focusing on **recall** can help identify customers who might leave our platform or service, allowing for more proactive measures to retain them.

Bussines Problem

ANALYTIC APPROACH

Here are the steps of the analysis which will undertake:

Step-1: Conduct Exploratory Data Analysis (EDA).
Step-2: Build a classification model based on behavioral analysis.
Step-3: Identify the factors that contribute to the likelihood of customers Churning.
Step-4: Develop a simulated scheme/strategy.

The analysis results can be accessed by stakeholders through a dedicated platform (Web/Mobile) whenever they need to perform retention activities.

Bussines Problem

METRIC EVALUATION

Confusion Matrix

Actual \ Predicted	0	1
0	TN customer actually does not churn and is predicted not to churn	FP customer actually does not churn and is predicted to churn
1	FN customer actually churns and is predicted not to churn	TP customer actually churns and is predicted to churn

Cost of FN (False Negative):

Disadvantages:
* Loss of a customer (churn) costing $100 per customer
* Cost of customer acquisition to replace the churned customer at $500 per customer

Emphasize **False Negatives** but also not forget about False Positives, with a greater focus on recall. Hence, the focus metric I use is the **F2-Score**.

CONTENTS

Data Understanding

- Dataset from the year 2022
- This dataset consists of 3941 rows and 11 columns

Table illustrating the data types along with their corresponding columns:

Type	Category	Columns
Numerical	Discrete	NumberOfDeviceRegistered, NumberOfAddress
	Continuous	Tenure, WarehouseToHome, DaySinceLastOrder, CashbackAmount
Categorical	Nominal	PreferedOrderCat, Marital_Status,
	Nominal(Binary)	Complain, Churn

Data Understanding

How does the data relate to the business context?

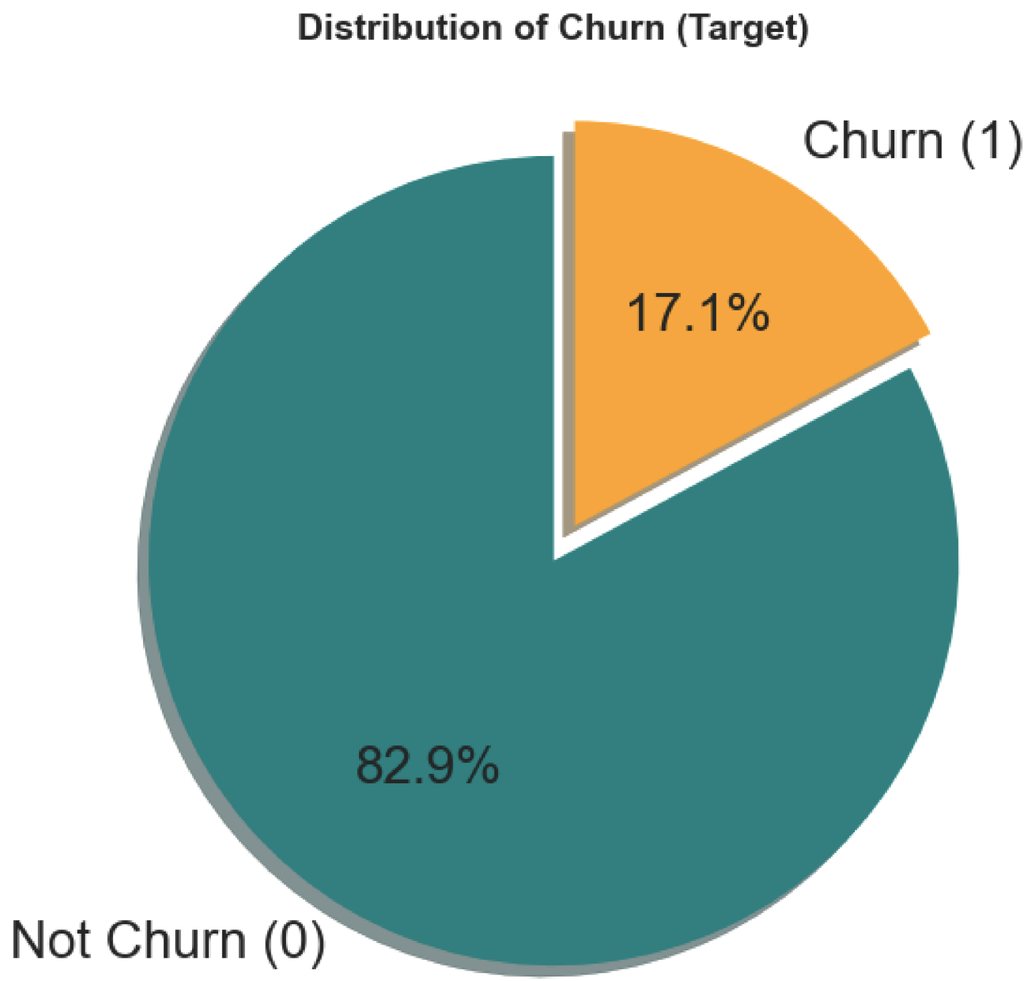
Based on range of quantitative and qualitative attributes across all variables and for all items:

	count	mean	std	min	25%	50%	75%	max
Tenure	3747.0	10.081398	8.498864	0.0	2.0	9.00	16.00	61.00
WarehouseToHome	3772.0	15.650583	8.452301	5.0	9.0	14.00	21.00	127.00
NumberOfDeviceRegistered	3941.0	3.679269	1.013938	1.0	3.0	4.00	4.00	6.00
SatisfactionScore	3941.0	3.088302	1.381832	1.0	2.0	3.00	4.00	5.00
NumberOfAddress	3941.0	4.237757	2.626699	1.0	2.0	3.00	6.00	22.00
Complain	3941.0	0.282416	0.450232	0.0	0.0	0.00	1.00	1.00
DaySinceLastOrder	3728.0	4.531652	3.667648	0.0	2.0	3.00	7.00	46.00
CashbackAmount	3941.0	176.707419	48.791784	0.0	145.7	163.34	195.25	324.99
Churn	3941.0	0.171023	0.376576	0.0	0.0	0.00	0.00	1.00

	PreferedOrderCat	MaritalStatus
count	3941	3941
unique	6	3
top	Laptop & Accessory	Married
freq	1458	2055

- Marital status: 52% are married
- Favorite product: 37% bought laptops and accessories
- Tenure status: 13% have used the app for 1 month
- Complaint status: 71% of customers did not complain
- Last purchase status: 16% made a purchase within the last 3 days
- Churn status: 83% of customers did not churn

Data Understanding



- 17.1% of customers have churned from this ecommerce service.

CONTENTS

Data Preparation

- **Data Cleaning**
  1. Handling Duplicate → 671 (17%) data, drop the same data.
  2. Handling Incosistent Data.

In the PreferredOrderCat column, there are inconsistent data, have the same meaning as other data. 'Mobile' and 'Mobile Phone' have the same meaning.

→ 'Mobile' is changed to 'Mobile Phone'.

Data Preparation

Split Dataset:

df_seen, df_unseen = train_test_split(df, test_size=0.2, random_state=RANDOM_STATE)

RANDOM_STATE=42

- **Data Preprocessing and Feature Engineering**

1. Imputation Missing Value
2. Binning Outliers
3. Encoding

CONTENTS

Modeling

Best Model

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	F2	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.8934	0.8837	0.5296	0.7414	0.6159	0.5561	0.5678	0.5607	0.3760
rf	Random Forest Classifier	0.8853	0.8878	0.4261	0.7650	0.5441	0.4848	0.5136	0.4662	0.4340
nb	Naive Bayes	0.7476	0.8082	0.7162	0.3654	0.4814	0.3392	0.3738	0.5978	0.1640

Naive Bayes, while not performing as well in most metrics, excels in recall and F2 score, making it useful in scenarios where identifying all positive cases is crucial in this case.

Modeling

Hyperparameter Tuning

- **Before Tuning**

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	F2
0	0.7327	0.7949	0.6471	0.3333	0.4400	0.2873	0.3146	0.5446
1	0.7279	0.8124	0.7500	0.3446	0.4722	0.3213	0.3654	0.6071
2	0.7153	0.8152	0.7910	0.3354	0.4711	0.3175	0.3722	0.6221
3	0.8254	0.8338	0.7164	0.4706	0.5680	0.4644	0.4805	0.6486
4	0.7368	0.7845	0.6765	0.3433	0.4554	0.3056	0.3361	0.5665
Mean	0.7476	0.8082	0.7162	0.3654	0.4814	0.3392	0.3738	0.5978
Std	0.0395	0.0171	0.0512	0.0527	0.0449	0.0637	0.0573	0.0376

- **After Tuning**

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	F2
0	0.7566	0.7987	0.6324	0.3583	0.4574	0.3157	0.3368	0.5485
1	0.7494	0.8202	0.7353	0.3650	0.4878	0.3459	0.3831	0.6112
2	0.7416	0.8205	0.7910	0.3605	0.4953	0.3528	0.4020	0.6386
3	0.8325	0.8349	0.7313	0.4851	0.5833	0.4839	0.4998	0.6640
4	0.7656	0.7928	0.6765	0.3770	0.4842	0.3480	0.3729	0.5838
Mean	0.7691	0.8134	0.7133	0.3892	0.5016	0.3693	0.3989	0.6092
Std	0.0327	0.0155	0.0543	0.0484	0.0428	0.0588	0.0547	0.0405

- **After Tuning** shows improved performance in most metrics.
- The overall improvements in **accuracy, precision, balanced F1 and F2 scores** suggest that **the tuned model performs better** than the untuned model.

CONTENTS

Evaluation Model

Interpretasi Model

How does Naive Bayes work?

Naive Bayes is a simple yet effective machine learning algorithm often used for classification. This algorithm is based on Bayes' Theorem with a strong (naive) assumption of independence between features.

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)}$$

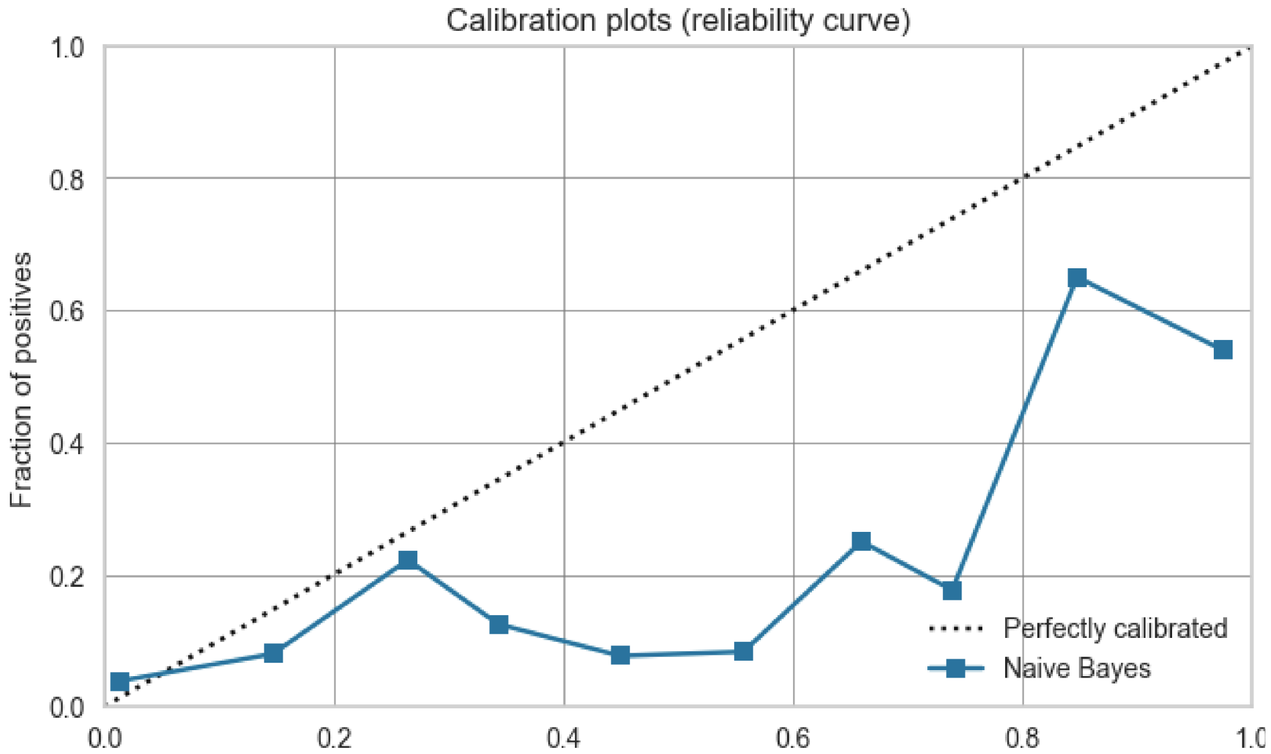
where:

- $P(C|X)$ is the posterior probability of class C given feature X.
- $P(X|C)$ is the likelihood probability of feature X given class C.
- $P(C)$ is the prior probability of class C.
- $P(X)$ is the total probability of feature X.

Evaluation Model

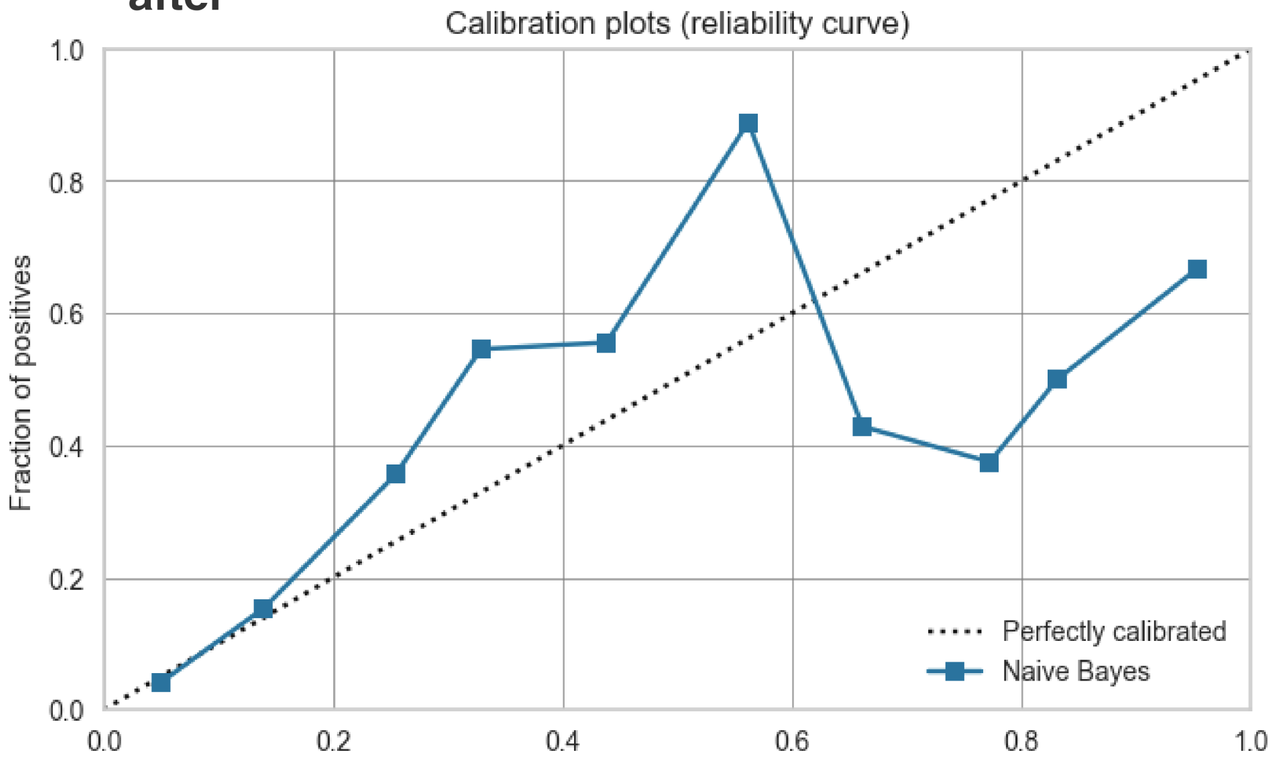
Calibration Model

before



- In the probability prediction range of 0.4 to 0.6, the model appears overly optimistic.

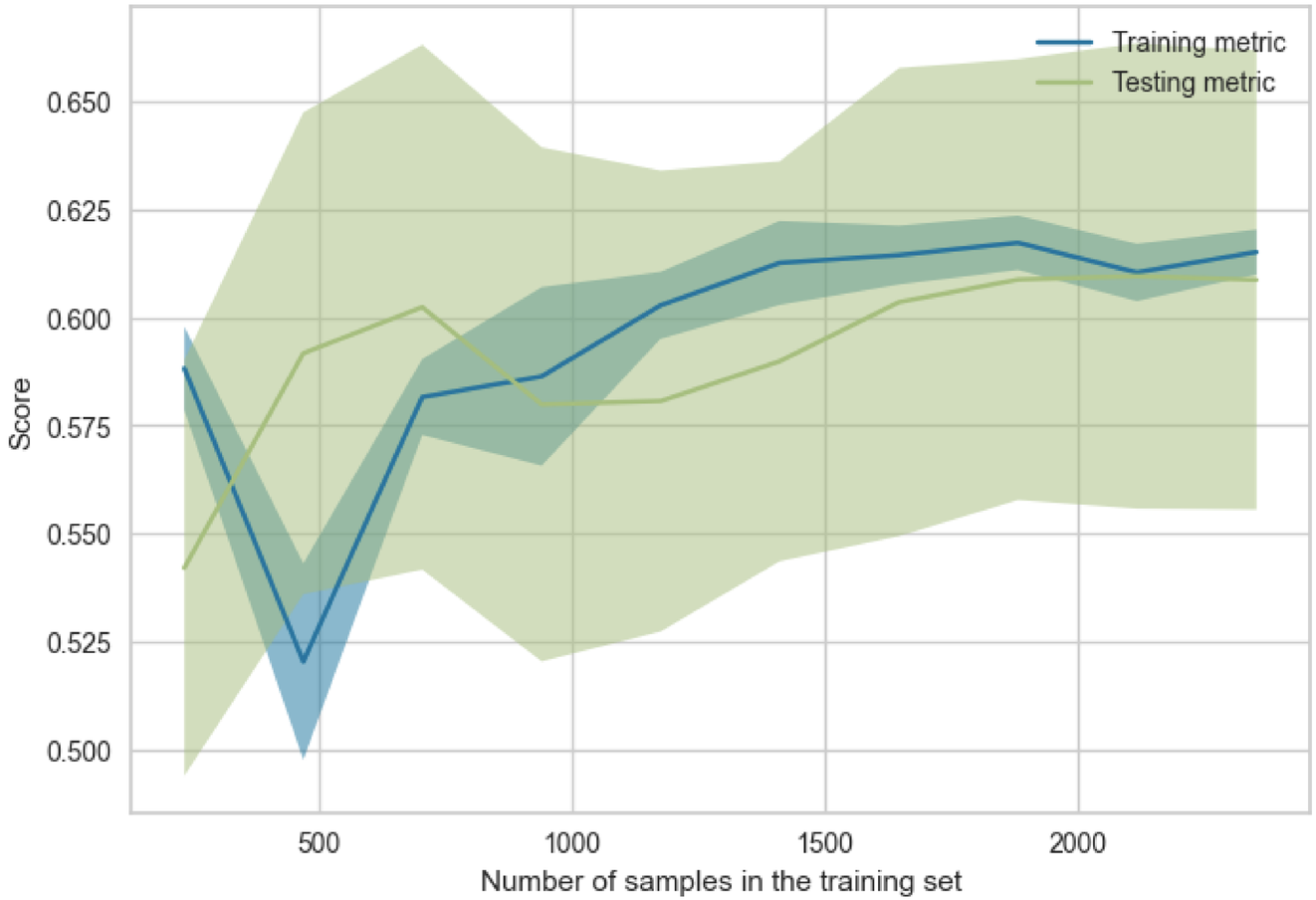
after



- In the probability prediction range of 0.5 to 0.6, the model appears overly optimistic.
- Conversely, in the prediction probability range of about 0.2 to 0.4 and 0.8 to 1.0, the model seems better calibrated or even slightly pessimistic.

Evaluation Model

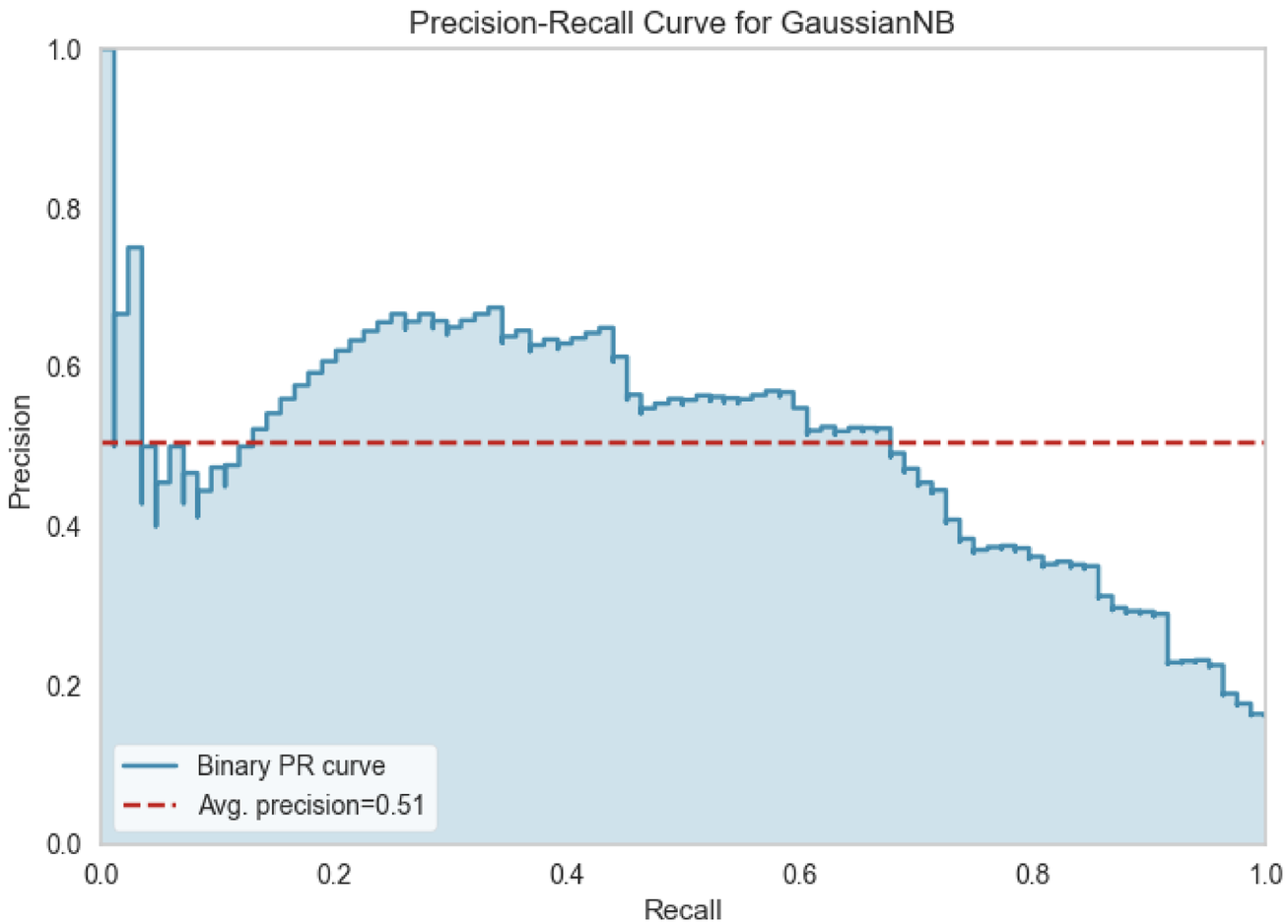
Check Overfitting or Underfitting



- Initial Overfitting
- Improvement with More Data
- The training and testing metrics converge and stabilize, indicating the model has reached its optimal performance given the data.
- Consistent Performance: With more data, the variability in the testing metric decreases, suggesting the model's performance becomes more reliable and consistent.

Evaluation Model

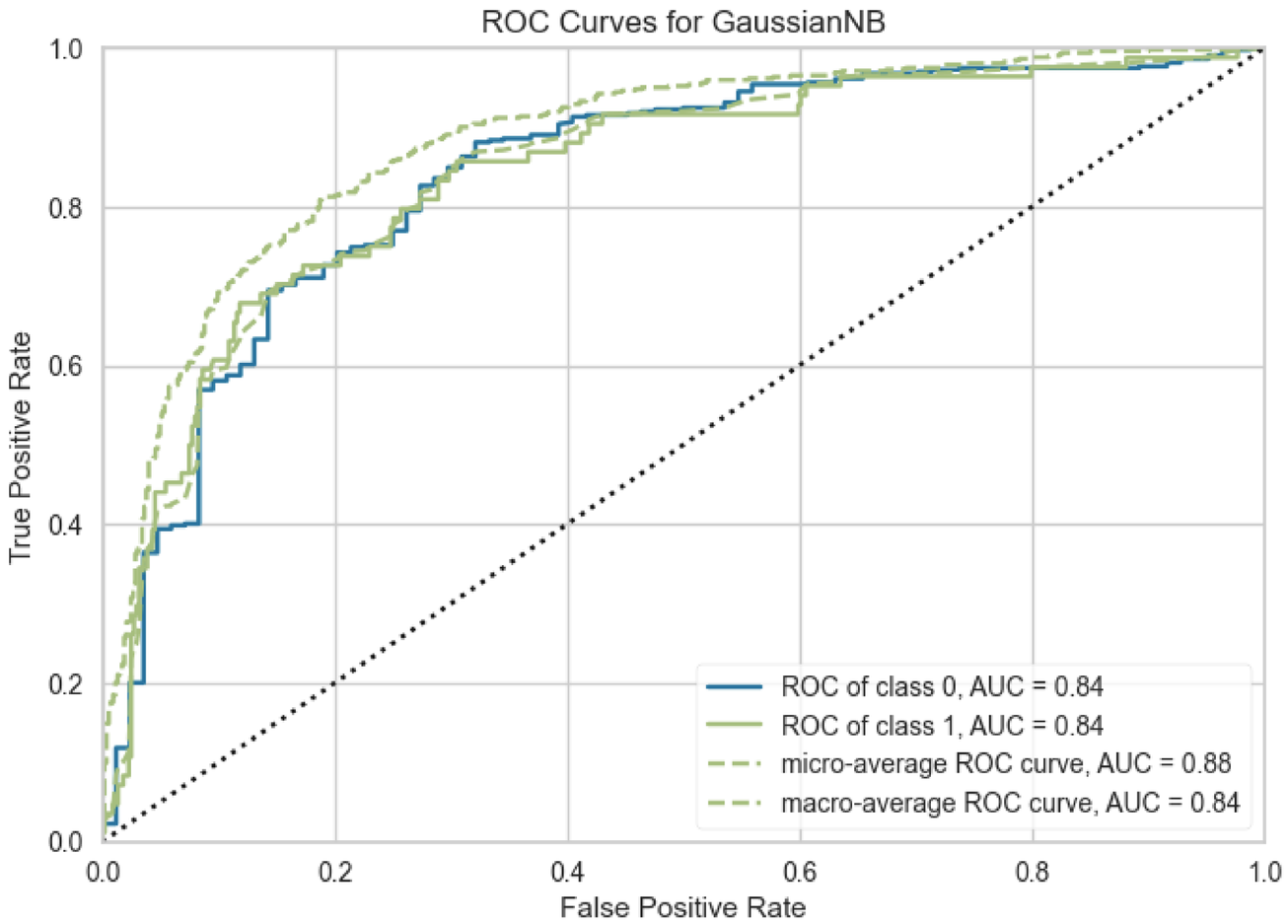
Check Precision-Recall Curve



- Indicates that the model has a moderate ability to balance precision and recall, with an average precision score of 0.51.
- The model shows variability in performance across different thresholds, and there are clear trade-offs between precision and recall.

Evaluation Model

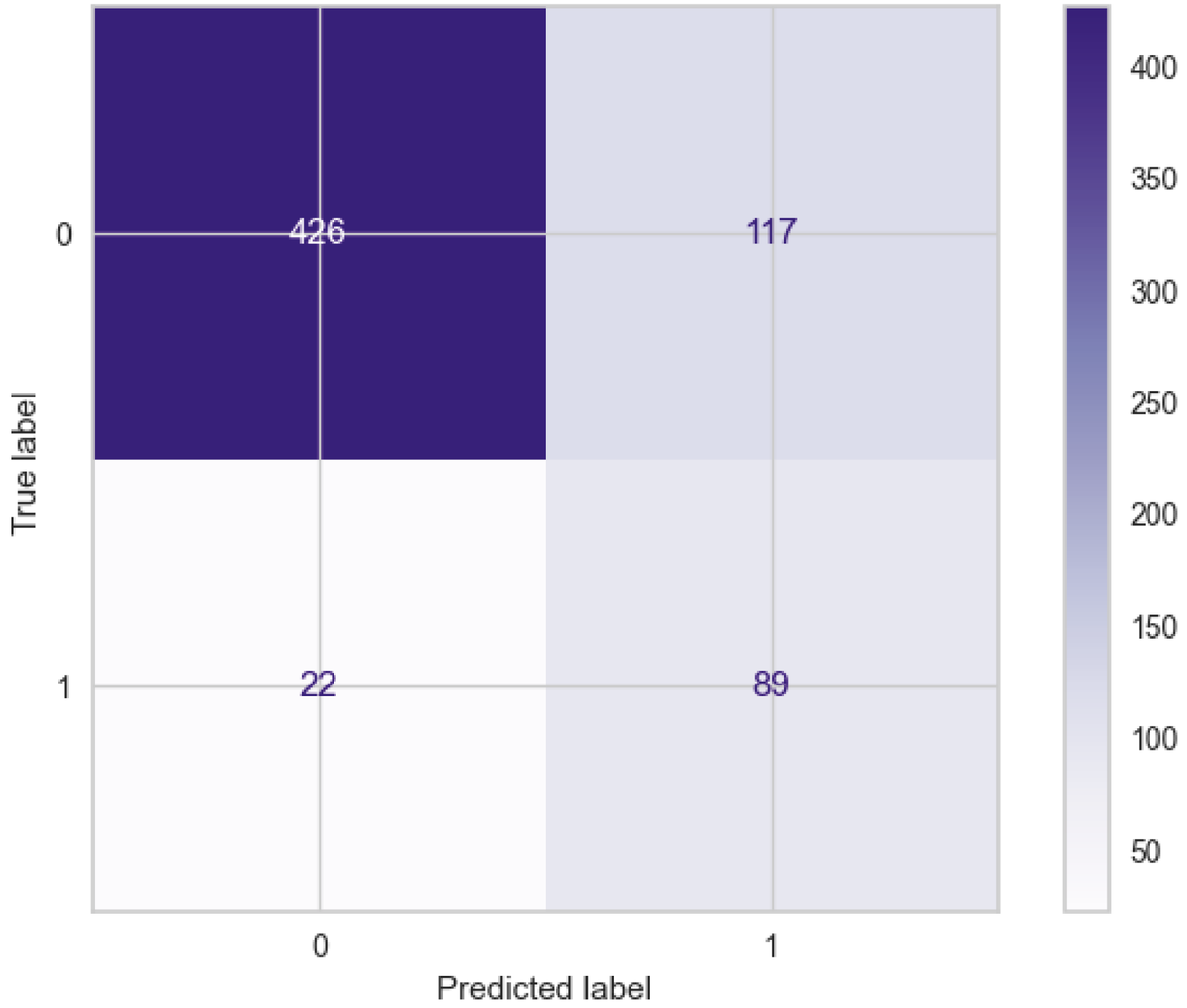
Check ROC Curve



- AUC Values: An AUC of 0.84 for both class 0 and class 1 indicates that the model has a good ability to distinguish between positive and negative classes.
- Micro vs. Macro Average: The AUC for the micro-average is 0.88, while the macro-average is 0.84. This suggests that the model performs slightly better when considering the proportion of instances in each class.

Evaluation Model

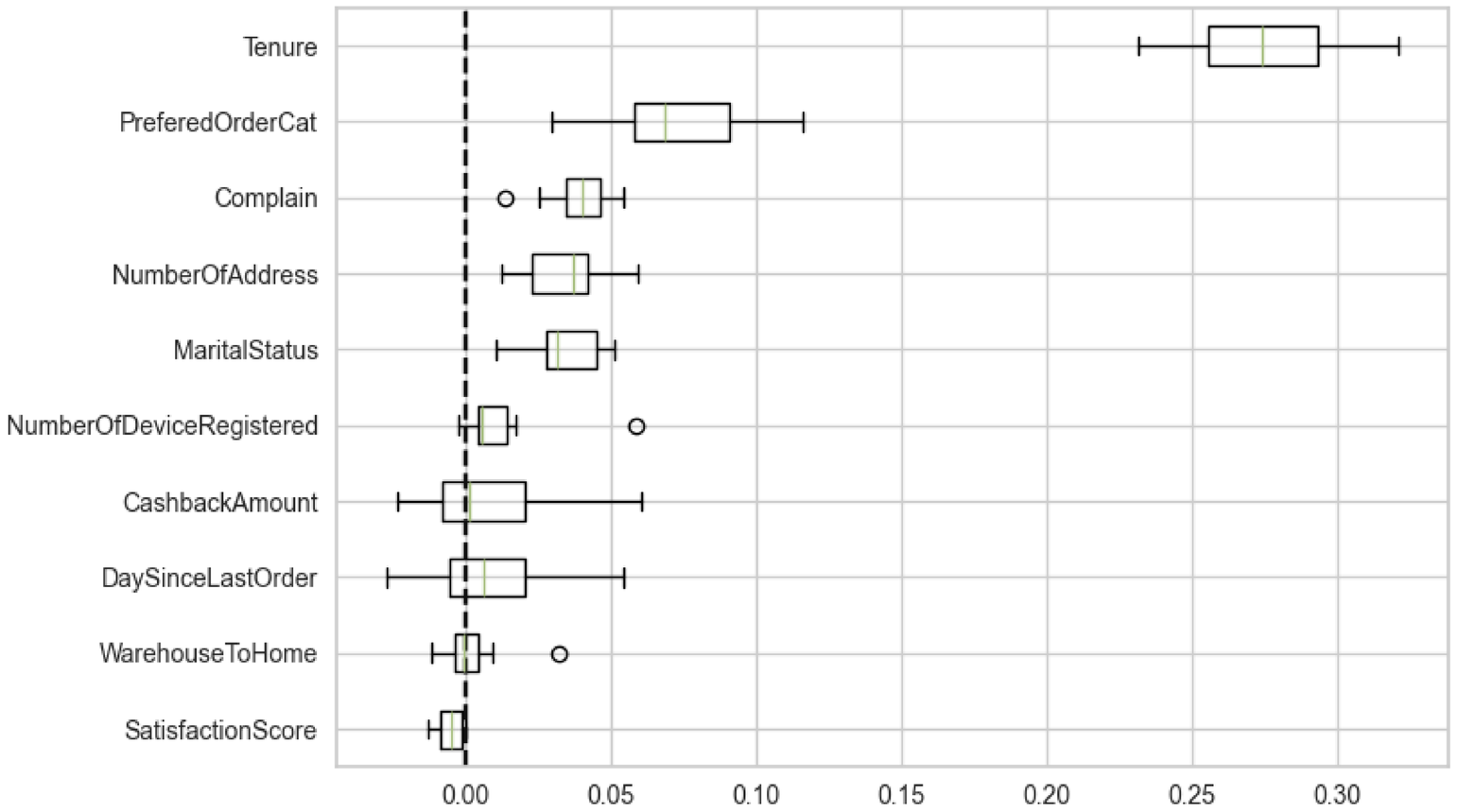
Confusion Matrix



- TP (True Positive): 89 (predicted that the customer will churn, and actually churned)
- TN (True Negative): 426 (predicted not to churn, and actually did not churn)
- FP (False Positive): 117 (predicted that the customer will churn, but actually did not churn)
- FN (False Negative): 22 (predicted that the customer will not churn, but actually churned)

Evaluation Model

Permutation Feature Importance



Interpretation of Each Feature:

- **Tenure**: This feature has a high importance score, with a median of around 0.29 and ranging from 0.25 to 0.31.
- **PreferredOrderCat**: This feature is also important, with a median of around 0.07
- **NumberOfAddress:**This feature has a median importance of around 0.05 with little variation.
- **Complain**: This feature has a median importance of around 0.04.
- **MaritalStatus:** This feature has a lower median, around 0.03.
- **DaySinceLastOrder, NumberOfDeviceRegistered, WarehouseToHome, CashbackAmount, SatisfactionScore:** All these features have low median importance scores, around 0.02 or less, with some variation in the distribution of values.

Evaluation Model

Counterfactuals

	Tenure	WarehouseToHome	NumberOfDeviceRegistered	PreferedOrderCat	SatisfactionScore	MaritalStatus	NumberOfAddress	Complain	DaySinceLastOrder	CashbackAmount	Churn
0	1.0	35.0	5	Laptop & Accessory	4	Married	2	0	7.0	165.490005	0

```
Diverse Counterfactual set (new outcome: 1)
```

	Tenure	WarehouseToHome	NumberOfDeviceRegistered	PreferedOrderCat	SatisfactionScore	MaritalStatus	NumberOfAddress	Complain	DaySinceLastOrder	CashbackAmount	Churn
2783	-	-	3	-	-	-	-	-	4.0	-	1
3699	-	-	-	Mobile Phone	-	Single	-	-	5.0	159.7899932861328	1
3108	-	31.0	-	-	-	Single	-	1	11.0	170.99000549316406	1

The counterfactual instances suggest that certain changes in features like **"PreferredOrderCat", "MaritalStatus", "DaySinceLastOrder", and "CashbackAmount"** can lead to a change in the churn outcome for the customer.

So, what's the correlation beetween feature importance with Tenure as the most important feature in the model?

This indicates that for some customers, churn prevention actions might be more effective if focused on changes in these features rather than solely on the duration of the customer's subscription.

CONTENTS

Conclusion

Without the model, it is difficult to know which customers will churn or not. Thus, the calculation is:

- Total customers: 654
- Acquisition cost to replace churned customers:
  654 * 500 USD = 327000 USD
  Total Cost: 327000 USD
  Amount saved: 0 USD

This means the potential acquisition costs incurred become much higher.

With the model:

Based on the confusion matrix:

- Cost for promotions: (117+89) * 100 USD = 20600 USD
- Acquisition cost to replace churned customers: 22 * 500 USD = 11000 USD
- **Total Cost: 20600 USD + 11000 USD = 31600 USD**
- **Amount saved: 327000 USD - 31600 USD = 295400 USD**

It is better to spend on retention costs rather than risking losing customers (**Churn**).

Recommendation

For Modeling:

- Ensure the data includes a customer ID column to avoid duplicate entries and no missing values.
- Add new features/columns that are related to the potential for customers to churn.
- Perform cohort analysis to observe churn based on customer transaction periods.
- Increase the sample size in the dataset to provide the model with more references.
- Experiment with different machine learning algorithms and conduct hyperparameter tuning.
- Select features according to their importance and try combinations of impactful features.

For Business:

- Create more segmented marketing campaigns
- Review the UI/UX (layout design) of the e-commerce application.
- Offer more promotions and excellent customer support initially to new customers.
- Consider providing more cashback, particularly to mobile phone buyers.
- Offer shipping cost discounts to potential churn customers who have a long distance.
- Pay attention to customer complaints regarding the ease of using the e-commerce platform.

– Thank You –

Do you have any question?
hilarykusuma@gmail.com