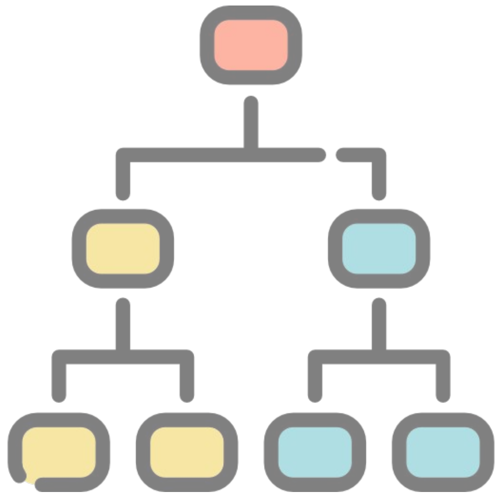
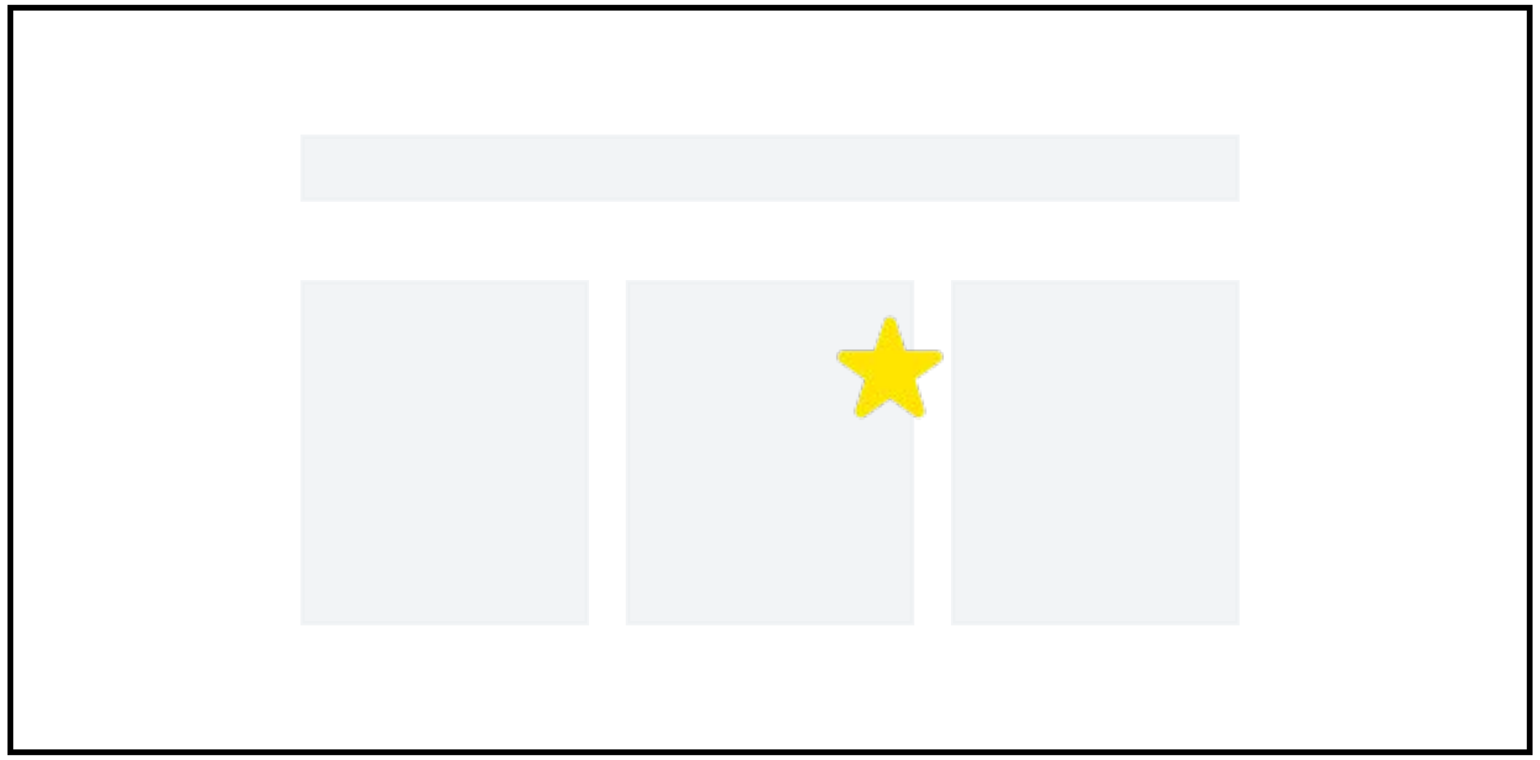
2022

Customer Churn Analysis and Prediction





The Team Members:
Anisa Ramadhani, S.Stat.
Ghifari, S.Kom.
Nurul Syiffa Yulianti, S.T.

Charlie Inc.

Charlie Inc.

Outline

2022

1. Business Problem Understanding
2. Data Understanding
3. Initial Data Analysis and Data Cleaning
4. Test Statistic
5. Exploratory Data Analysis (EDA)
6. Data Splitting and Data Preprocessing
7. Model Benchmarking and Modeling on Test Data
8. Hyperparameter Tuning
9. Feature Importance
10. Conclusion
11. Recommendation

Business Problem Understanding

Role: E-Commerce Consultant

1. Churned customers may cause company loss due to missing potential orders and purchases
2. Mistargeted Investments by the company may also decrease profits and increase losses

Our Recommendation : Churn Prediction

- Classification Target 1 = Churn
- Classification Target 0 = Not Churn

1. What factors influence customer Churn behavior?

2. What business strategy is more suitable to minimize customer churn?

1. Identify the factors that may cause customer churn and predict the customer churn potential

2. Analyze business strategies to minimize losses by using vouchers to minimize customer churn.

1 Context

2 Problem Statement

3 Goals

Method: Supervised Machine Learning ->Classification

4 Analytic Approach

F1 Score

5 Metric Evaluation

Assumptions:

- Churned customers are the customers who delete their account.
- The E-Commerce presumably located in India because UPI (Unified Payments Interface) in the dataset may refer to payment service that integrates the services of banks in India on a single application.

Data Understanding

Attribute	Data Type, Length	Description	
CustomerID	Integer	Unique customer ID	
Churn	Integer	Binary Target: 0 - Not churn, 1 - Churn	
Tenure	Float	Tenure of customer in organization	Month
PreferredLoginDevice	String	Preferred login device of customer	
CityTier*	Integer	City tier	
WarehouseToHome	Float	Distance in between warehouse to home of customer	Kilometer
PreferredPaymentMode	String	Preferred payment method of customer	
Gender	String	Gender of customer	
HourSpendOnApp	Float	Number of hours spend on mobile application or website	(Daily Average)
NumberOfDeviceRegistered	Integer	Total number of deceives is registered on particular customer	
PreferedOrderCat	String	Preferred order category of customer in last month	
SatisfactionScore*	Integer	Satisfactory score of customer on service	
MaritalStatus	String	Marital status of customer	
NumberOfAddress	Integer	Total number of added added on particular customer	
Complain*	Integer	Binary, 1 - If any complaint has been raised in last month	
OrderAmountHikeFromlastYear	Float	Percentage increases in order from last year	
CouponUsed	Float	Total number of coupon has been used in last month	
OrderCount	Float	Total number of orders has been places in last month	
DaySinceLastOrder	Float	Day Since last order by customer	
CashbackAmount	Integer	Average cashback in last month	INR

Assumptions:

1. Some features are categorical (Nominal, Ordinal, Binary), with somewhat low cardinality. The highest cardinality contains 7 unique values

2. There are some redundant unique values where 2 values has similar meaning, we can merge them into a single value to reduce the cardinality

3. Every row represents exactly 1 customer, whether it be churned ones or still active customers

* Categorical features which don't need to be encoded

Data Understanding

2

Attribute	Data Type, Length	Description
CustomerID	Integer	Unique customer ID
Churn	Integer	Binary Target: 0 - Not churn, 1 - Churn
Tenure	Float	Tenure of customer in organization
PreferredLoginDevice	String	Preferred login device of customer
CityTier*	Integer	City tier
WarehouseToHome	Float	Distance in between warehouse to home of customer
PreferredPaymentMode	String	Preferred payment method of customer
Gender	String	Gender of customer
HourSpendOnApp	Float	Number of hours spend on mobile application or website
NumberOfDeviceRegistered	Integer	Total number of deceives is registered on particular customer
PreferedOrderCat	String	Preferred order category of customer in last month
SatisfactionScore*	Integer	Satisfactory score of customer on service
MaritalStatus	String	Marital status of customer
NumberOfAddress	Integer	Total number of added added on particular customer
Complain*	Integer	Binary, 1 - If any complaint has been raised in last month
OrderAmountHikeFromlastYear	Float	Percentage increases in order from last year
CouponUsed	Float	Total number of coupon has been used in last month
OrderCount	Float	Total number of orders has been places in last month
DaySinceLastOrder	Float	Day Since last order by customer
CashbackAmount	Integer	Average cashback in last month

Total data: 5630 rows, 20 columns

Data types :

- float (7 columns),
- integer (8 columns),
- object (5 columns).

Categorical Features

Numerical Features

Target

* Categorical features which don't need to be encoded

Data Understanding

Categorical Descriptive Statistic

	count	unique	top	freq
PreferredLoginDevice	5630	3	Mobile Phone	2765
PreferredPaymentMode	5630	7	Debit Card	2314
Gender	5630	2	Male	3384
PreferedOrderCat	5630	6	Laptop & Accessory	2050
MaritalStatus *	5630	3	Married	2986
Complain *	5630	2	1	4026
CityTier *	5630	3	3	3666
SatisfactionScore	5630	5	5	1698

- Feature with **highest cardinality** (7 unique values) : PreferredPaymentMode
- Features with **lowest cardinality** (2 unique values/binary) : Gender, Complain

Numerical Descriptive Statistic

	count	mean	std	min	25%	50%	75%	max
CustomerID	5630.0	52815.500000	1625.385339	50001.0	51408.25	52815.5	54222.75	55630.0
Churn	5630.0	0.168384	0.374240	0.0	0.00	0.0	0.00	1.0
Tenure *	5366.0	10.189899	8.557241	0.0	2.00	9.0	16.00	61.0
CityTier	5630.0	1.654707	0.915389	1.0	1.00	1.0	3.00	3.0
WarehouseToHome	5379.0	15.639896	8.531475	5.0	9.00	14.0	20.00	127.0
HourSpendOnApp	5375.0	2.931535	0.721926	0.0	2.00	3.0	3.00	5.0
NumberOfDeviceRegistered *	5630.0	3.688988	1.023999	1.0	3.00	4.0	4.00	6.0
SatisfactionScore	5630.0	3.066785	1.380194	1.0	2.00	3.0	4.00	5.0
NumberOfAddress *	5630.0	4.214032	2.583586	1.0	2.00	3.0	6.00	22.0
Complain	5630.0	0.284902	0.451408	0.0	0.00	0.0	1.00	1.0
OrderAmountHikeFromlastYear	5365.0	15.707922	3.675485	11.0	13.00	15.0	18.00	26.0
CouponUsed	5374.0	1.751023	1.894621	0.0	1.00	1.0	2.00	16.0
OrderCount	5372.0	3.008004	2.939680	1.0	1.00	2.0	3.00	16.0
DaySinceLastOrder	5323.0	4.543491	3.654433	0.0	2.00	3.0	7.00	46.0
CashbackAmount	5630.0	177.221492	49.193869	0.0	146.00	163.0	196.00	325.0

Max Tenure = 61 (month)
Max HourSpendOnApp = 5 (daily average)

* Categorical features which don't need to be encoded

Initial Data Analysis and Data Cleaning

Delete unique ID; Delete duplicated rows; Delete category redundancy;

Drop Unique Identifier

\- - - - -

Unique identifier will not be used for analysis and modeling

- 1 column dropped: 'CustomerID'

Drop Duplicated Rows

\- - - - -

Duplicated rows may caused overfit or information leakage

- 556 duplicated rows deleted

Merge Redundant Values

\- - - - -

Merge similar values to a single value

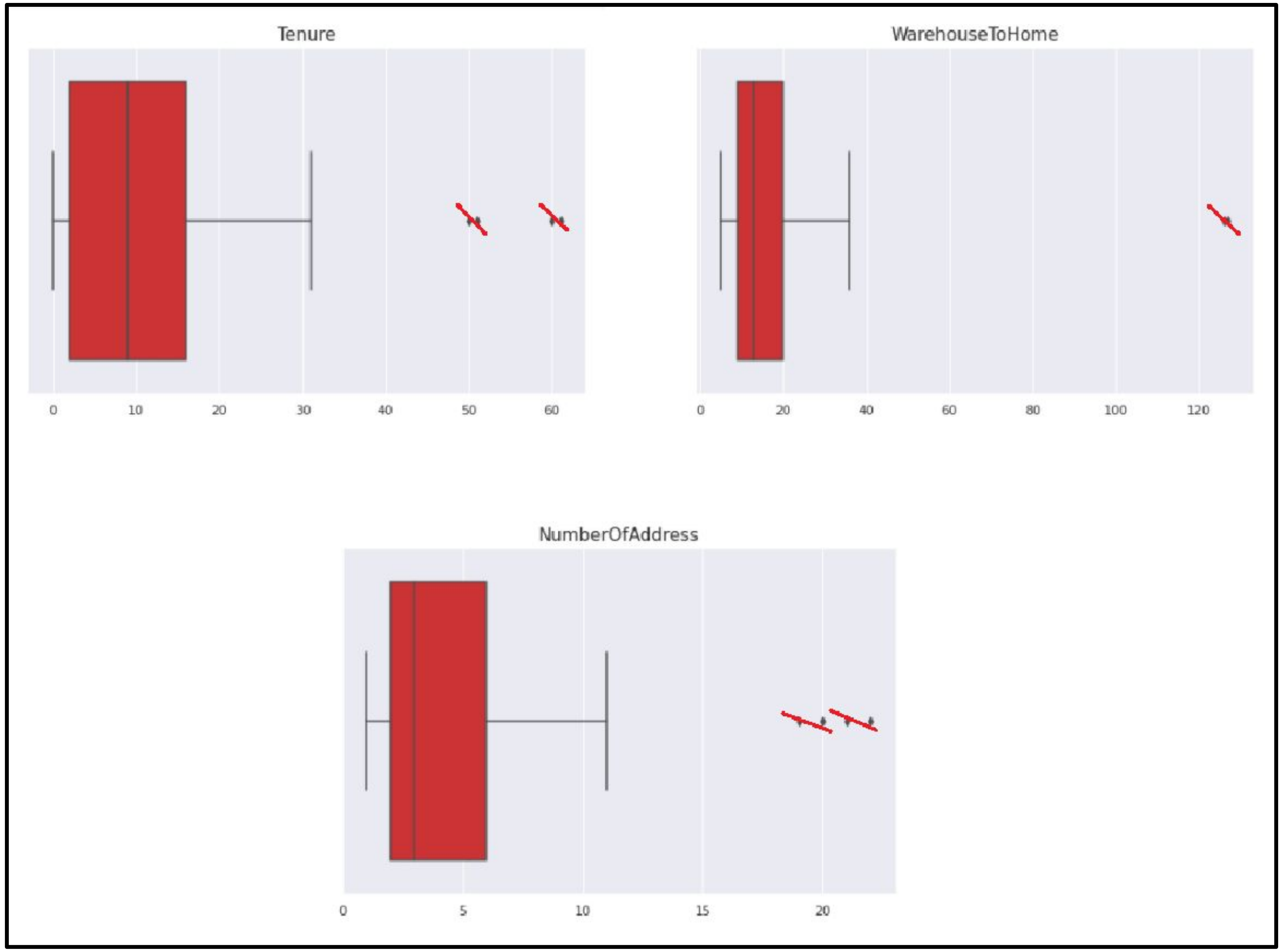
- Mobile, Phone & Mobile Phone
- CC & Credit Card
- COD & Cash on Delivery

Clean Data
=
Good Model

Initial Data Analysis and Data Cleaning

3

Handling Outliers



1. 4 data points have tenure far longer than the others, but they only have very few orders (3 max), also their last order is also not too long ago.

2. 2 data points have home very far away from the nearest warehouse. Due to the far distance and the very few amounts, these outliers can be deleted.

3. 4 data points had input many addresses on their app, however they only did 2 orders maximum.

These outliers are few and very far from the rest of the observations, so we can delete these outliers in particular..

Total data:

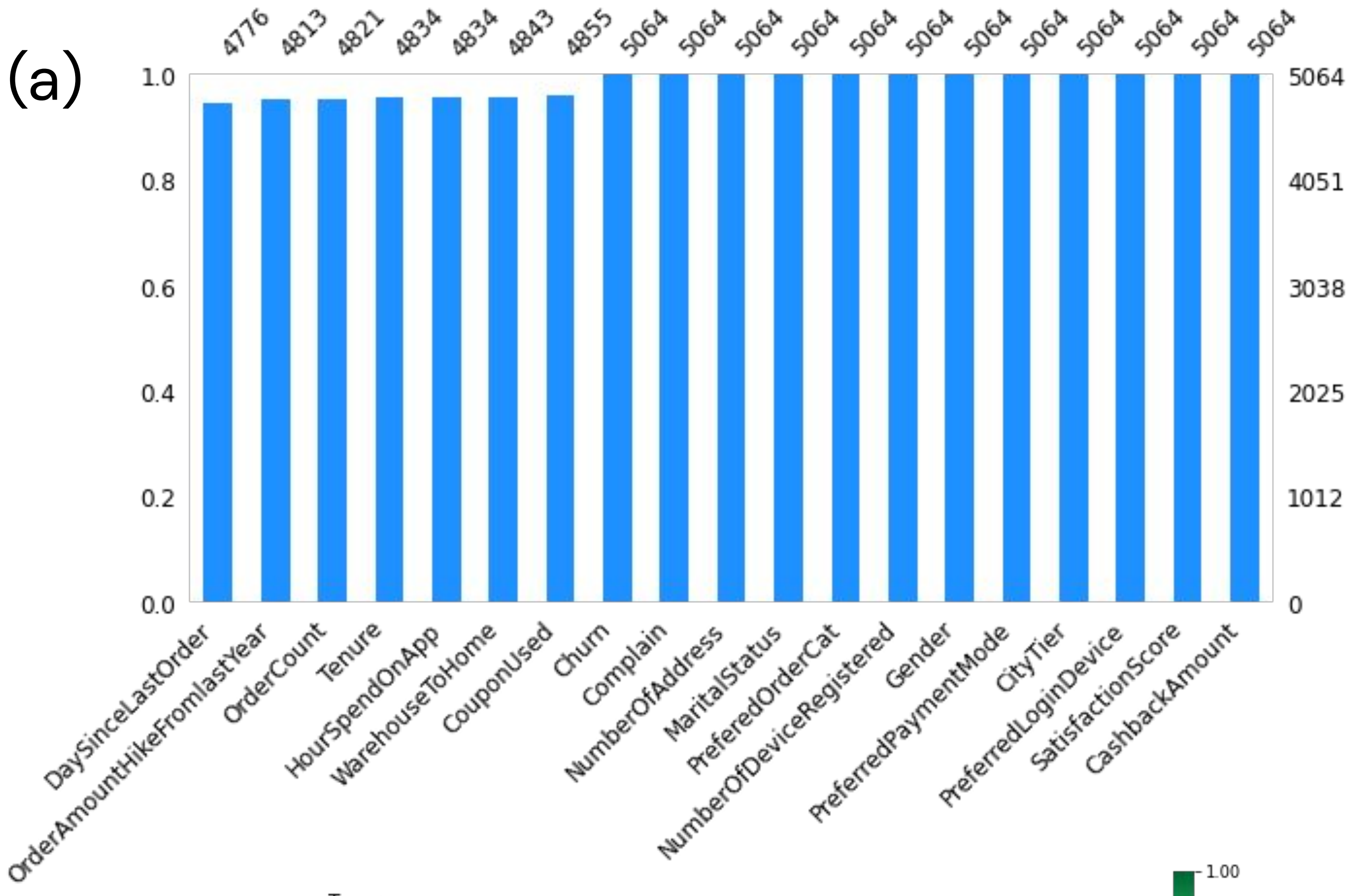
- **Before** dropping duplicates and outliers: **5630 rows, 20 columns**
- **After** dropping duplicates and outliers: **5064 rows, 19 columns**

Initial Data Analysis and Data Cleaning

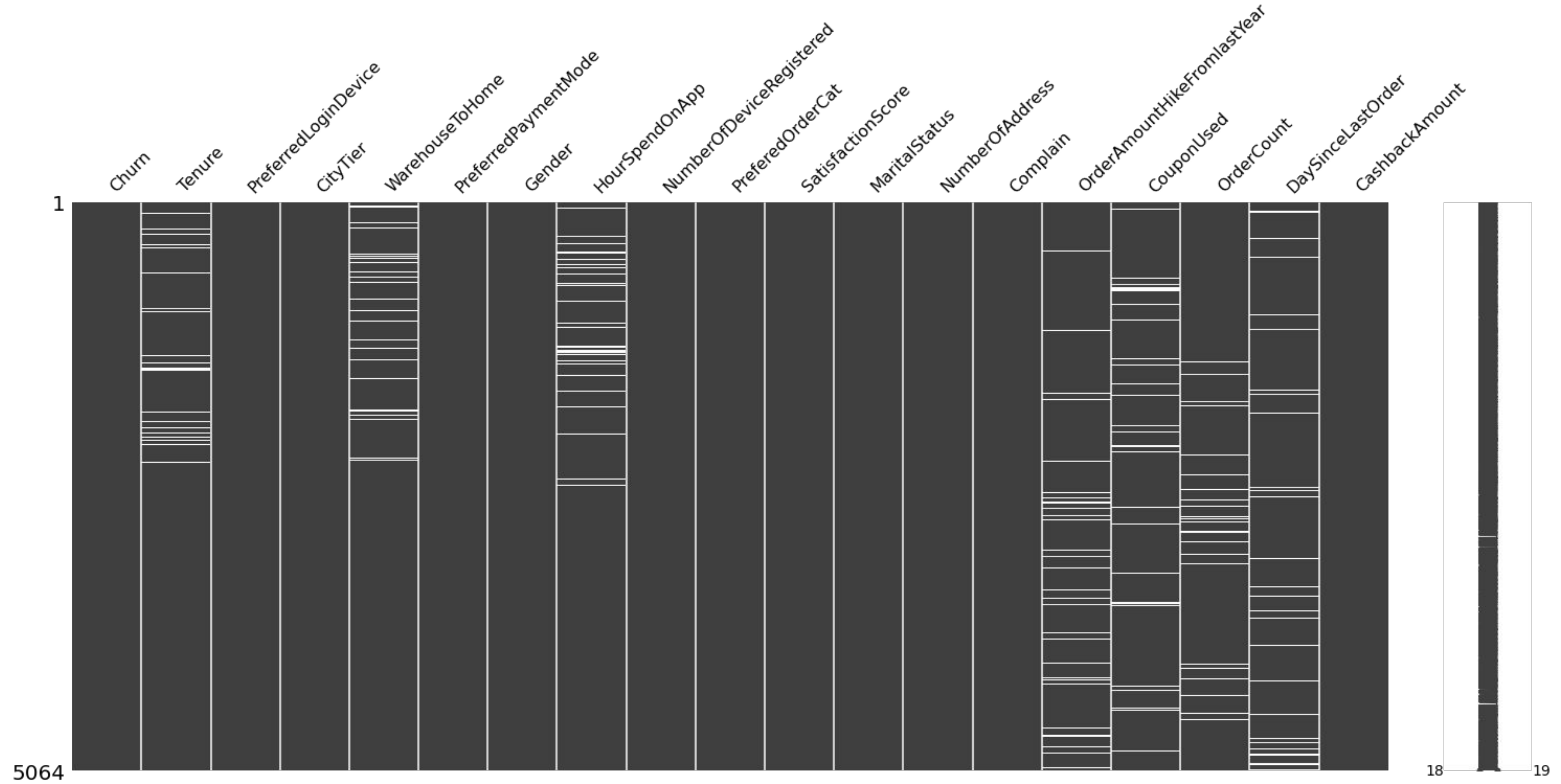
3

Identify Missing Values

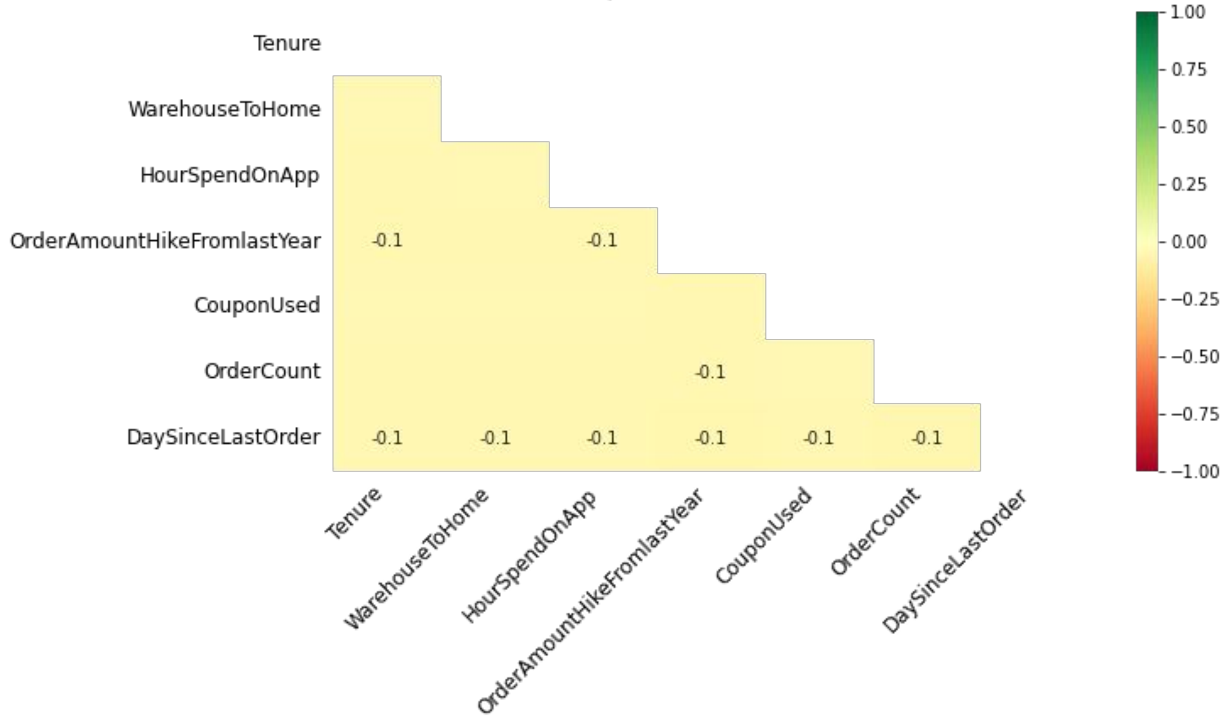
(a)



(b)



(c)



Missing Value : DaySinceLastOrder, OrderAmountHikeFromLatyear, OrderCount,Tenure, HourSpendOnApp, WarehouseToHome, CouponUsed (a)

Missing values' pattern are random (b) and not correlated one another (c).

Initial Data Analysis and Data Cleaning

Handling Missing Values

	Tenure	Warehouse ToHome	HourSpend OnApp	NumberOf Device Registered	NumberOf Address	OrderAmount HikeFromlast Year	Coupon Used	OrderCount	DaySince LastOrder	Cashback Amount
Tenure		-0.0447	-0.0057	-0.0222	0.2889	0.0207	0.1236	0.1736	0.2054	0.4457
Warehouse ToHome	-0.0447		0.0825	0.0253	0.0228	0.0402	0.0184	0.0331	0.0368	0.0233
HourSpend OnApp	-0.0057	0.0825		0.3535	0.1834	0.1360	0.3119	0.2901	0.1093	0.1890
NumberOf Device Registered	-0.0222	0.0253	0.3535		0.1270	0.0964	0.2525	0.2501	0.0359	0.2094
NumberOf Address	0.2889	0.0228	0.1834	0.1270		0.0446	0.1052	0.0697	-0.0775	0.2631
OrderAmount HikeFromlast Year	0.0207	0.0402	0.1360	0.0964	0.0446		0.0678	0.0600	0.0064	0.0361
CouponUsed	0.1236	0.0184	0.3119	0.2525	0.1052	0.0678		0.7246	0.3291	0.3263
OrderCount	0.1736	0.0331	0.2901	0.2501	0.0697	0.0600	0.7246		0.4778	0.3908
DaySince LastOrder	0.2054	0.0368	0.1093	0.0359	-0.0775	0.0064	0.3291	0.4778		0.3822
Cashback Amount	0.4457	0.0233	0.1890	0.2094	0.2631	0.0361	0.3263	0.3908	0.3822	

0.7246

'CouponUsed' & 'OrderCount' are the only features that are strongly correlated (0.72)

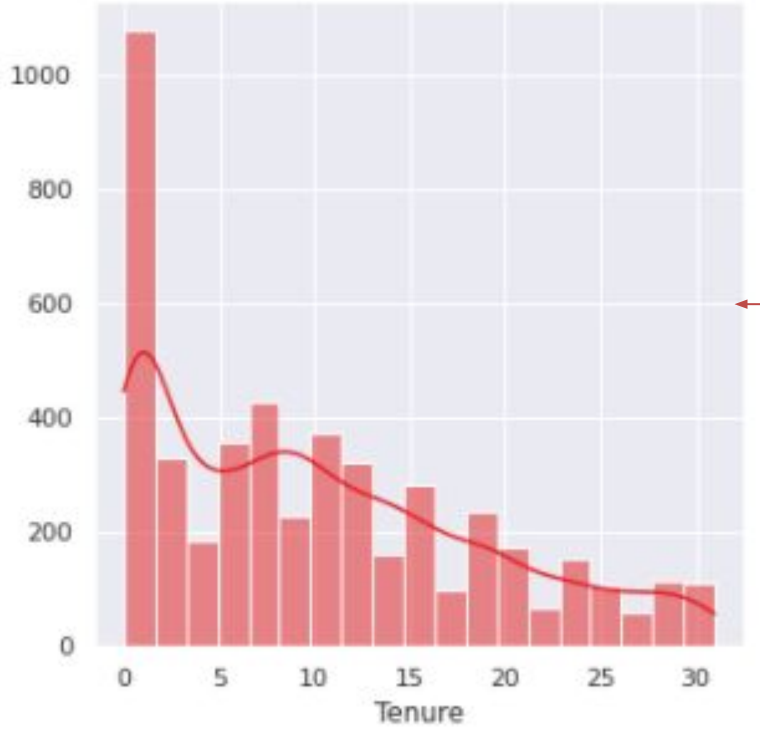
-> both will be **iteratively imputed**

Initial Data Analysis and Data Cleaning

Handling Missing Values

Normality Test

	feature	normal_stats	p_val	hypothesis
0	WarehouseToHome	0.887448	2.382207e-44	Not Normal (Reject H0)
1	Tenure	0.906713	1.884186e-41	Not Normal (Reject H0)
2	DaySinceLastOrder	0.890055	5.465064e-44	Not Normal (Reject H0)
3	OrderAmountHikeFromlastYear	0.915125	4.998586e-40	Not Normal (Reject H0)
4	HourSpendOnApp	0.827627	0.000000e+00	Not Normal (Reject H0)



The rest of the features containing missing values have non-normal distribution, including:

1. 'DaySinceLastOrder'
2. 'WarehouseToHome'
3. 'Tenure'
4. 'OrderAmountHikeFromlastYear'
5. 'HourSpendOnApp'

-> they all will be **imputed** with their respective **'median'** values

After being imputed, there are no more missing values

Test Statistic

4

Two Sample Independent T-Test

Hypothesis:

Ho: The average value of non churn **equals** the one that is churn from 1 feature

Ha: The average value of non churn does **not equal** the one that is churn from 1 feature

	feature	t_stats	p_val	hypothesis
0	Tenure	-25.154585	5.721674e-132	Dependent (Reject H0)
1	WarehouseToHome	4.988401	3.145865e-07	Dependent (Reject H0)
2	HourSpendOnApp	0.970007	1.660446e-01	Independent (Accept H0)
3	NumberOfDeviceRegistered	8.363530	3.893265e-17	Dependent (Reject H0)
4	NumberOfAddress	3.470546	2.618686e-04	Dependent (Reject H0)
5	OrderAmountHikeFromlastYear	-1.511510	6.536053e-02	Independent (Accept H0)
6	CouponUsed	-0.192087	4.238409e-01	Independent (Accept H0)
7	OrderCount	-1.542998	6.144683e-02	Independent (Accept H0)
8	DaySinceLastOrder	-10.587844	3.161856e-26	Dependent (Reject H0)
9	CashbackAmount	-10.246951	1.055610e-24	Dependent (Reject H0)

Conclusion: P-Value is lower than the significance level (0.05) = we have sufficient evidence to reject the **Ho**.
>> numeric features that have different mean between churn and not churn:
Tenure, WarehouseToHome, NumberOfDeviceRegistered, NumberOfAddress, D aySinceLastOrder, dan **CashbackAmount.**

Chi-Square

Hypothesis:

Ho: Feature does not affect the target (churn)

Ha: Feature affects the target (churn)

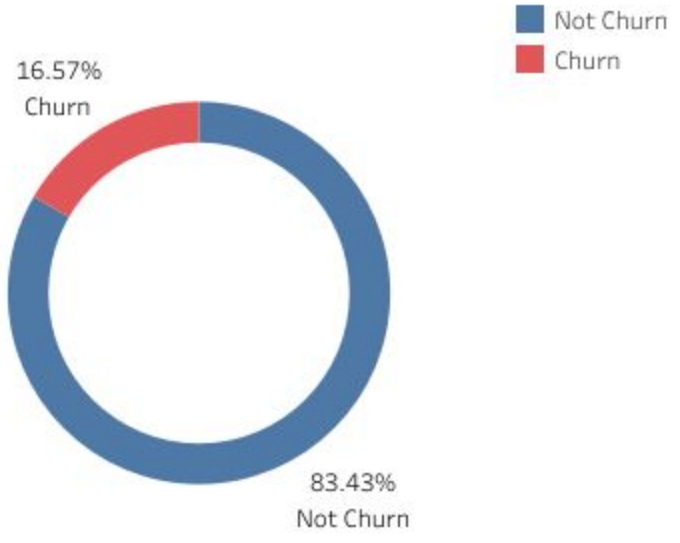
	feature	chi_stats	p_val	hypothesis
0	PreferredLoginDevice	12.385324	4.327215e-04	Dependent (Reject H0)
1	CityTier	49.952380	1.422259e-11	Dependent (Reject H0)
2	PreferredPaymentMode	49.755702	4.060830e-10	Dependent (Reject H0)
3	Gender	3.973949	4.620927e-02	Dependent (Reject H0)
4	PreferedOrderCat	226.426179	7.760609e-48	Dependent (Reject H0)
5	SatisfactionScore	57.479912	9.811513e-12	Dependent (Reject H0)
6	MaritalStatus	169.697011	1.415019e-37	Dependent (Reject H0)
7	Complain	309.645877	2.608364e-69	Dependent (Reject H0)

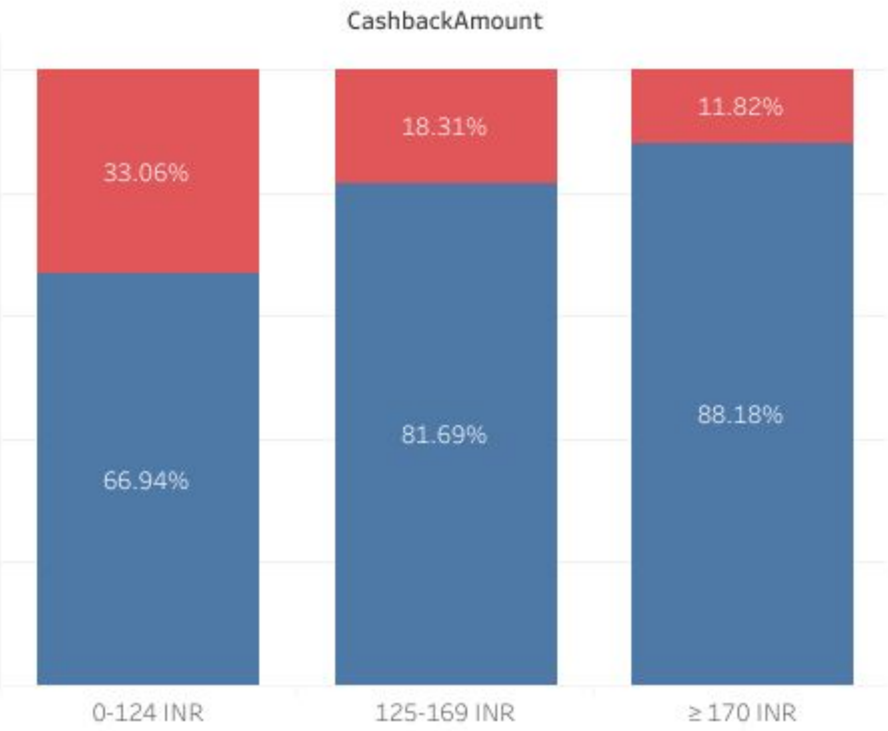
Conclusion: P-Value is lower than the significance level (0.05) = we have sufficient evidence to reject the **Ho**.
>> **all** categorical features affect 'Churn'.

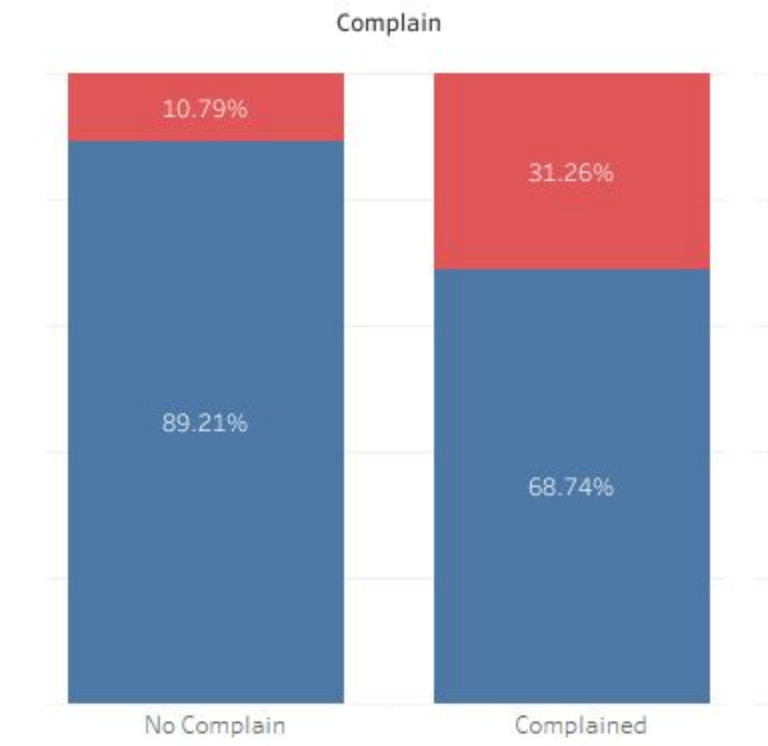
Exploratory Data Analysis

Where's the E-Commerce Located?

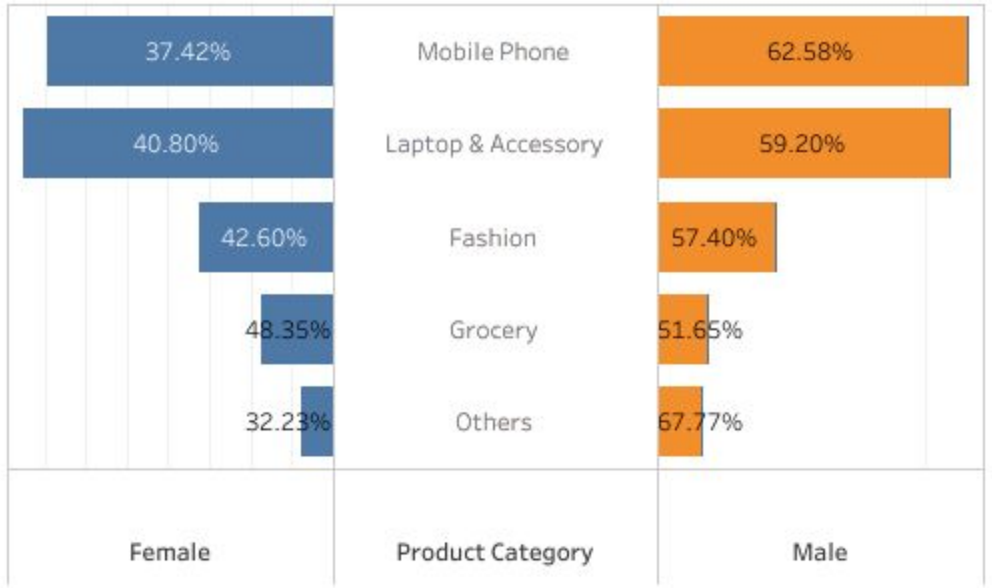
How many customers have churned?



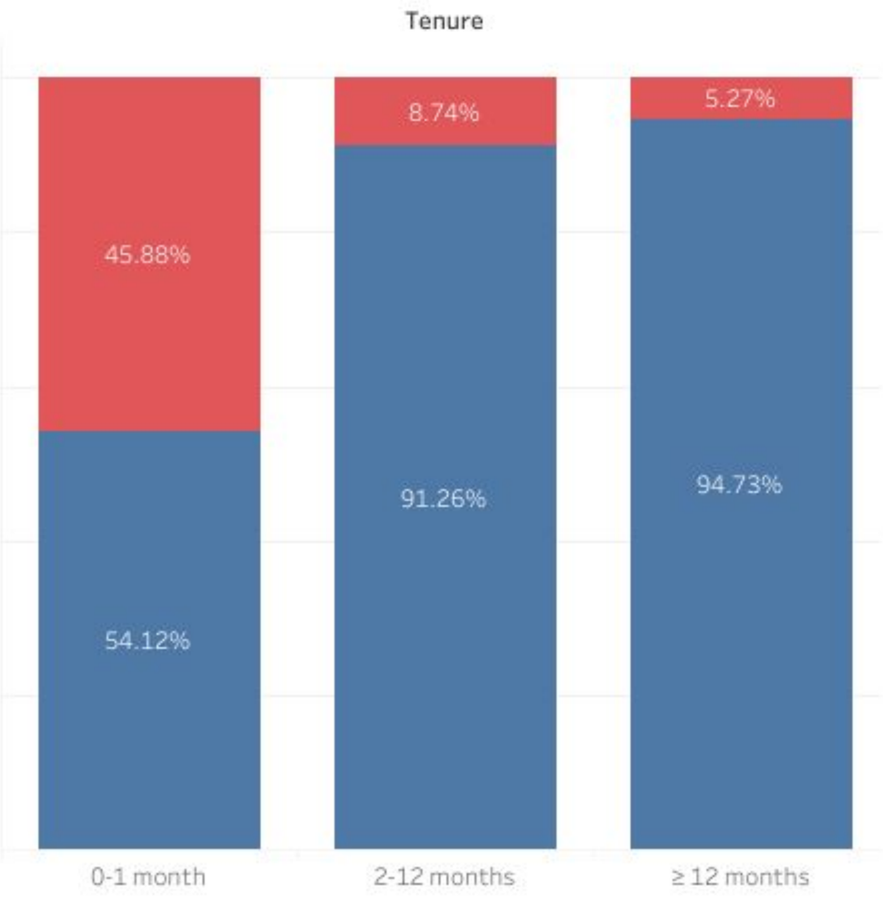


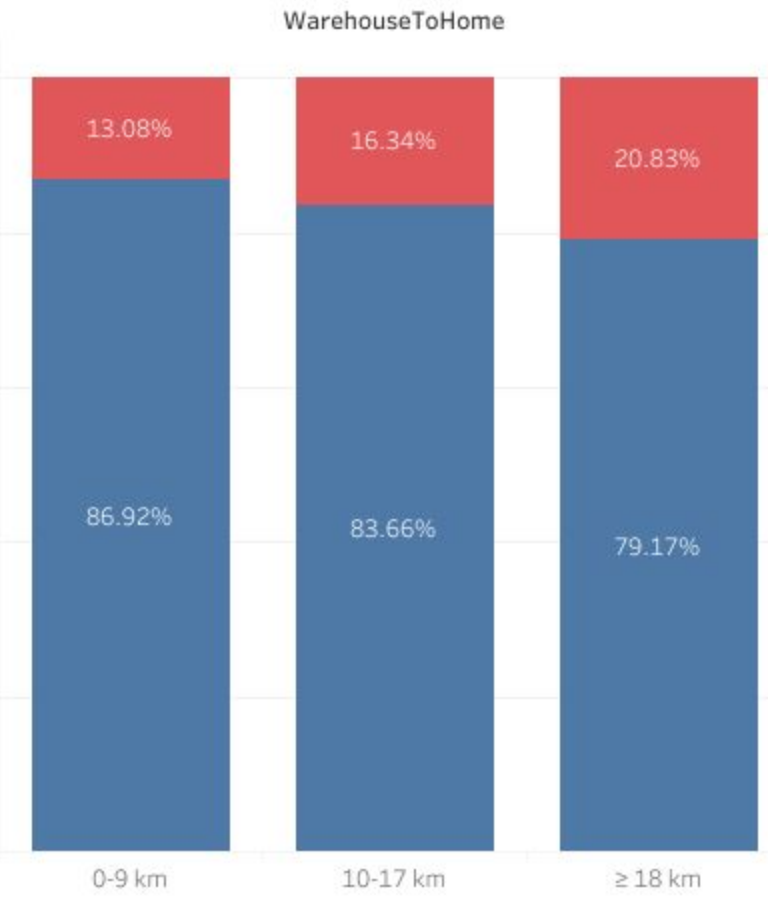


Item category purchased based on gender:



- More customers purchased 'mobile phone' and 'laptop & accessory' than the other categories.
- Male customers purchased more items than female customers in general.
- Customers are more likely to churn when they have lower cashback and tenure.
- Customers are more likely to churn when they have reported a complaint.
- Customers are more likely to churn when they have further distance from their home to the nearest warehouse.





Created in Tableau; Charlie Inc. 2022

Data Splitting & Preprocessing

Data Splitting

TRAIN SET : TEST SET

80% : 20%

Data Preprocessing

One Hot Encoding

(≤4 unique values)

- PreferredLoginDevice
- Gender
- MaritalStatus

Binary Encoding

(>4 unique values)

- PreferredPaymentMode
- PreferredOrderCat

Modelling

Model Benchmark: Train Set

Model	Mean F1-Score	Std
LightGBM	0.874574	0.025805
Random Forest	0.870420	0.029388
Decision Tree	0.820771	0.021130
Adaptive Boosting	0.820761	0.027525
Gradient Boosting	0.707085	0.052768
XGBoost	0.684257	0.044418
Logistic Regression	0.571417	0.067113
KNN	0.473248	0.048200

From the results above, **Random Forest** and **LightGBM** give a higher F1-score compared to other models

Test Set

Model	F1-Score
LightGBM	0.865031
Random Forest	0.867925

After fitting the model to test set, there is no big difference between the two models F1-score. Next, hyperparameter tuning will be performed on both models to improve their performance.

Hyperparameter Tuning

PARAMETER

LightGBM

- 'model__max_bin'
- 'model__num_leaves'
- 'model__min_data_in_leaf'
- 'model__num_iterations'
- 'model__learning_rate'

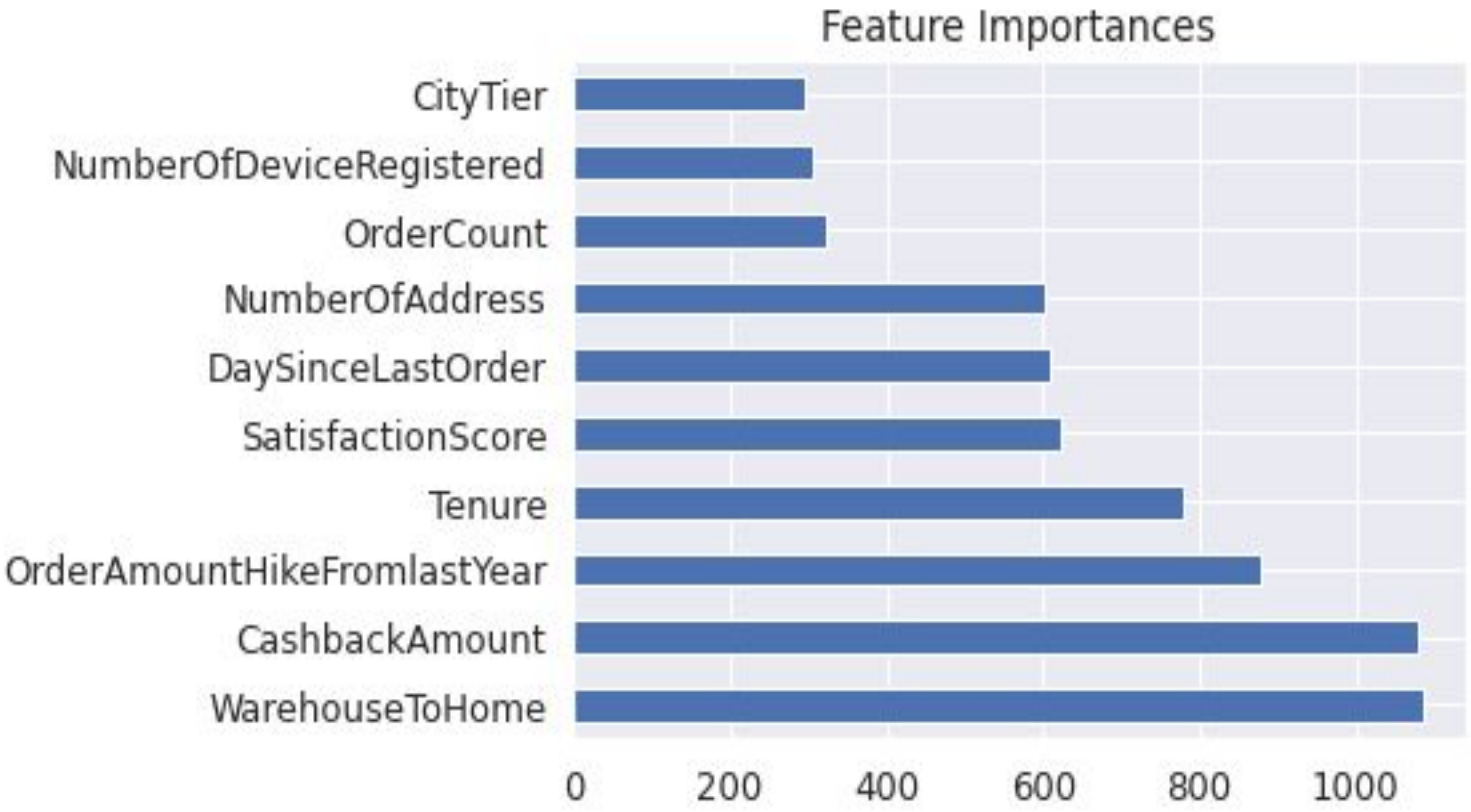
Random Forest

- 'model__n_estimators'
- 'model__criterion'
- 'model__max_features'
- 'model__max_depth'
- 'model__min_samples_split'
- 'model__min_samples_leaf'
- 'model__bootstrap'

Model	F1-Score Before Tuning		F1-Score After Tuning	
	Train Set	Test Set	Train Set	Test Set
LightGBM	0.874574	0.865031	0.910657	0.901493
Random Forest	0.870420	0.867325	0.886497	0.875379

Based on the results of hyperparameter tuning, the best model obtained is **LightGBM** because it has a higher F1-score.

Feature Importances

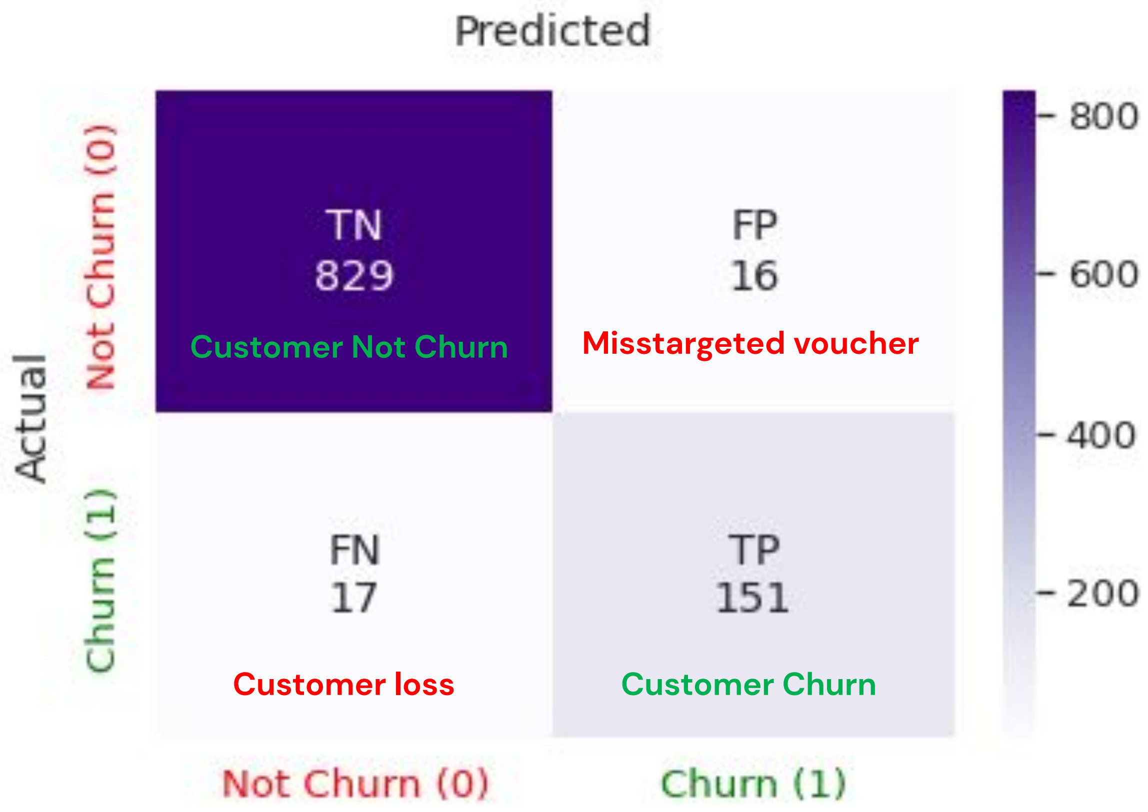


Based on the graph, it can be seen that the most important features are **‘WarehouseToHome’** and **‘CashbackAmount’**.

Conclusion

Classification	PRECISION	RECALL	F1-SCORE
1	0.90	0.90	0.90
0	0.98	0.98	0.98

1 = Churn ; 0 = Not Churn



Assumptions:

a. 1013 customers; 845 not churn; 168 churn.

b. Average Cashback Amount (not churn) = 180 INR per customer

c. Cost for voucher = 190 INR per customer.

d. Cost for voucher divided into class (per customer) = Low; 180 INR, Medium; 185 INR, High; 190 INR.

e. Loss due to customer churn : 2 order per customer * 500 INR = 1000 INR per customer.

Cost Benefit Analysis

Without model

Total cost for voucher:
190 INR * 1013 cust =
192.470 INR

Total loss (mistargeted voucher):
190 INR * 845 cust =
160.550 INR

Saves around **83.5%**

Prediction using LGBM

Total cost for voucher:
190 INR * (151+16) cust
31.730 INR

Total loss:
(190 INR * 16 cust) + (1000 INR * 17 cust)
20.040 INR

Saves around **0.3%**

Voucher given divided into class

Low Churn Probability:
- **Total Cost**: 6 cust * 180 INR = 1.080 INR
- **Total Loss**: 4 cust * 180 INR = 720 INR

Medium Churn Probability:
- **Total Cost**: 7 cust * 185 INR = 1.295 INR
- **Total Loss**: 4 cust * 185 INR = 740 INR

High Churn Probability:
- **Total Cost**: 154 cust * 190 INR = 29.260
- **Total Loss**: 8 cust * 190 INR = 1.520 INR

Total cost for voucher:
31.635 INR

Total loss:
2.980 INR + (17 cust * 1000 INR)
19.980 INR

Recommendation

1. Analyze the possible cause behind why is the test score often higher than the train score.

2. Add more data and gather more information to strengthen assumptions that have been made.

3. Add more features related to the amount customers spend on the E-Commerce. The added features may improve model performance and assist in determining strategies to reduce customer churn.

4. More research is needed regarding the current E-Commerce business strategy related to ‘WarehouseToHome’.

5. Try other Machine Learning algorithms and do hyperparameters tuning again to get better result.

6. More thorough analysis on the wrong prediction result.

7. Identify the direction of the feature's relationship to the target by using Shap library.

Thank you!