Data-Driven Insights for GTM, and Product

Javad Anaraki
2025.06.06

Data Exploration

- The input data has 20 features and 5203 samples. Out of 20 feature, Churn is the target with two values of 0 and 1, and CustomerID include ids.

	Churn	
Labels	0	1
Counts	4329	874

- Based on the counts, the input dataset is **unbalanced** so should be mindful of this.

Data Exploration (cont.)

- Numerical features
  - Churn, Tenure, CityTier, WarehouseToHome, HourSpendOnApp, NumberOfDeviceRegistered, SatisfactionScore, NumberOfAddress, Complain, OrderAmountHikeFromlastQuarter, CouponUsed, OrderCount, DaySinceLastOrder, CashbackAmount
- Non-numerical features
  - PreferredLoginDevice, PaymentMode, Gender, OrderCat, MaritalStatus

* Non-numerical features might need to be converted to **categorical features** or in some case to one-hot-enconding for specific models

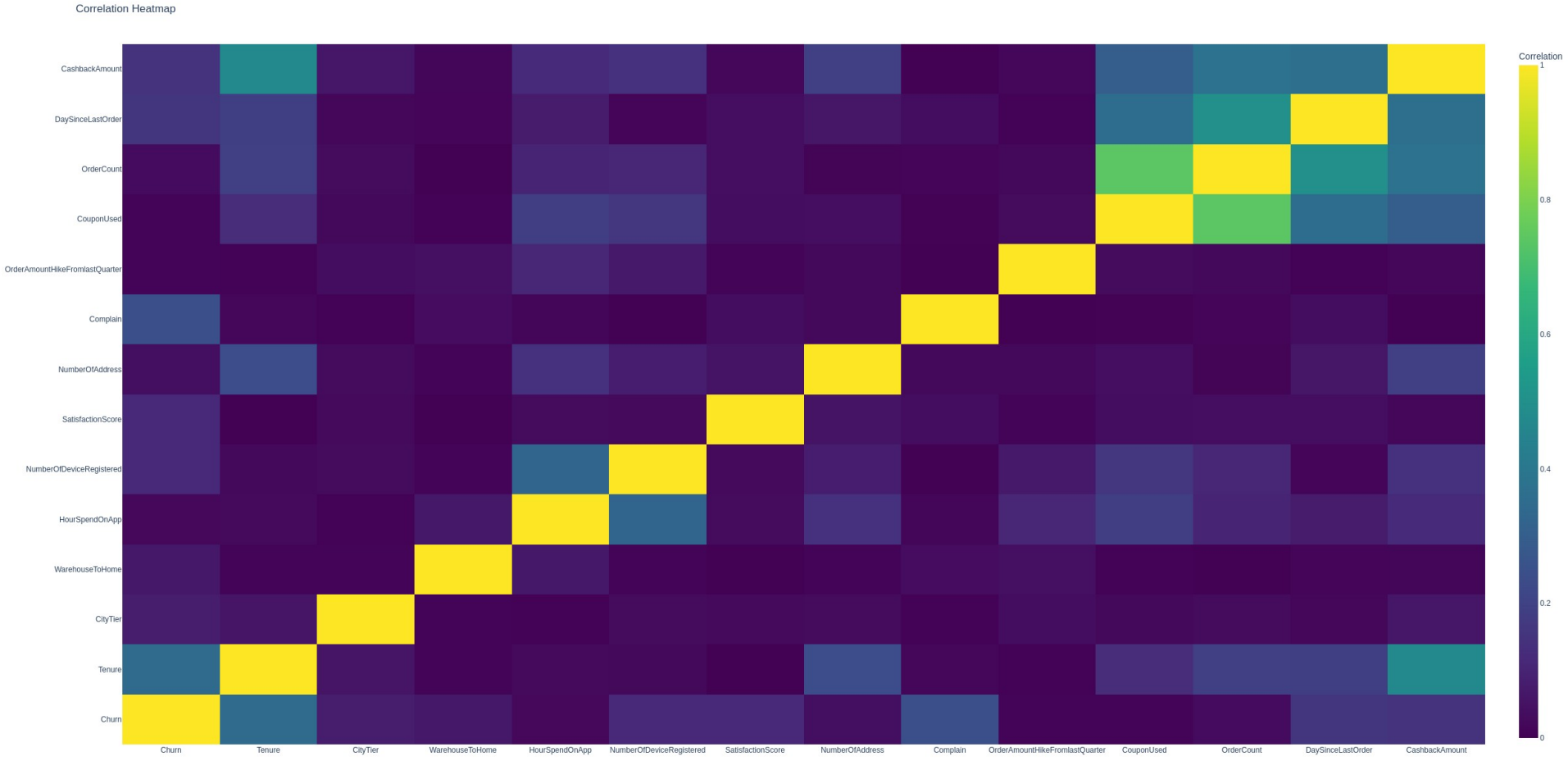
Data Exploration (cont.)

- Looking at the info() function results shows that the input dataset has **missing values**. Although there are models that can handle missing values but we would need to have an imputation strategy (for numerical and categorical) in case we need classification methods that can’t handle missing values.
  - Tenure (239)
  - WarehouseToHome (238)
  - HourSpendOnApp (229)
  - OrderAmountHikeFromlastQuarter (237)
  - CouponUsed (237)
  - OrderCount (251)
  - DaySinceLastOrder (292)

Data Exploration (cont.)

- In **PreferredLoginDevice,** there are three categories which two of them, namely, Mobile Phone and Phone should the same. Therefore, we need to handle this by setting the same value for both.
- In **PaymentMode**, there are six categories where we have the same category but different names, such as Credit Card and CC, and COD and Cash on Delivery, which needs to be handled.
- In **OrderCat**, Mobile Phone and Mobile can be merged.

Analysis from CX perspectives



Analysis from CX perspectives (cont.)

- Highly correlated features
  - OrderCount & CashbackAmount
  - OrderCount & DaySinceLastOrder
  - OrderCount & CouponUsed
  - Tenure & CashbackAmount
- Highly correlated feature to the outcome
  - Tenure
  - Complain
  - DaySinceLastOrder
  - CashbackAmount

Insights and Recommendations

- Generating model to predict churn probability would help us improve the CX and prevent customers leaving the company. To this goal, I trained to models, namely, RF and XGBoost. The reason for selecting these two models are:
  - RF
    - Robust churn prediction on structured data
    - Handles missing values
    - Reduces overfitting of single trees
  - XGBoost
    - Handles complex datasets with mixed types
    - Handles missing values
    - Handles class imbalance well

Insights and Recommendations (cont.)

- For the metrics, I used f1-score and confusion matrix. The reason to select these methods are:
- f1-score
  - Good metric for datasets with imbalance outcome
  - Useful when you want a balance between precision and recall
- Confusion matrix
  - Shows all types of classification outcomes (TP, TN, FP, FN)
  - Help further analyze the results

Insights and Recommendations (cont.)

- RF
  - f1-score: 0.845665961945031 7
  - TN, FP, FN, TP: 1288, 11, 62, 200
  - 123 customers with probability > 70% are likely to churn
- XGBoost
  - f1-score: 0.901353965183752 4
  - TN, FP, FN, TP: 1277, 22, 29, 233
  - 240 customers with probability > 70% are likely to churn

Insights and Recommendations (cont.)

- Customers with the median Tenure of 1.0 are likely to churn.
- Customers with the median Tenure of 10.0 are likely to stay.

- Customers with the median WarehouseToHome of 15.0 are likely to churn.
- Customers with the median WarehouseToHome of 13.0 are likely to stay.

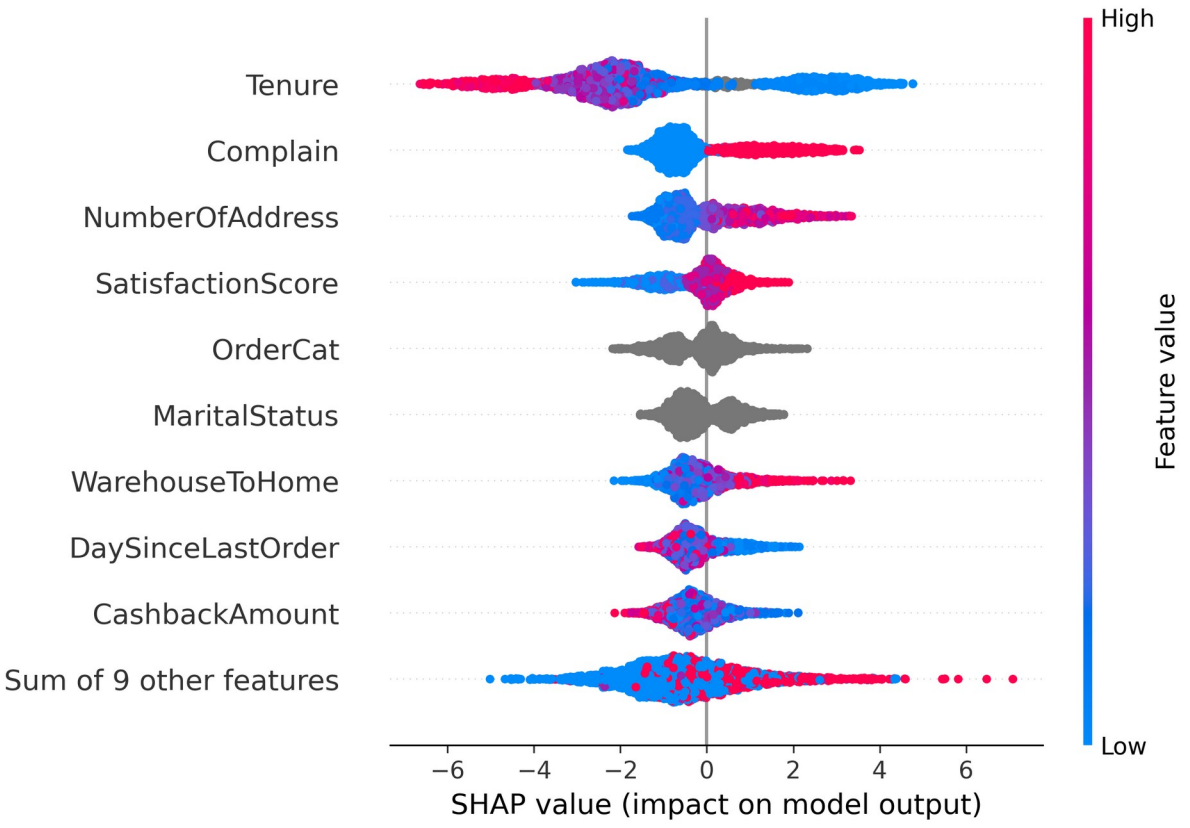
- Customers with the median OrderAmountHikeFromlastQuarter of 14.0 are likely to churn.
- Customers with the median OrderAmountHikeFromlastQuarter of 15.0 are likely to stay.

- Customers with the median CashbackAmount of 150.53500000000003 are likely to churn.
- Customers with the median CashbackAmount of 166.63 are likely to stay.

Insights and Recommendations (cont.)

- Providing higher Cashback and Coupon would positively impact customer churn
- Providing higher Cashback and Coupon improves DaySinceLastOrder which would decrease Complains
- Increasing warehouses to decrease WarehouseToHome would positively impact Complains and eventually help user’s retention

Analysis from CX perspectives (cont.)



Thank you!