Customer Churn

Analysis, Insights and Prediction

Project Overview

This project aims to:

- Analyze customer churn data
- Derive insights and recommendations for CX department
- Train Machine Learning model/s to predict churn

Please Note: The detailed analysis, observations and insights are captured in the Jupyter notebook uploaded on my public github: https://github.com/jhalalit/Churn_Insights_and_Prediction.git

Exploratory Data Analysis

Data variability, Skewness, Missing Values, Outliers

Data distribution analysis - Preprocessing

- **Features with Right-Skew:**

  – *Tenure, WarehouseToHome, CashbackAmount etc.*

  Handling: A log-transformation can help bring them closer to normality, useful for regression based models.

- **Features with Outliers:**

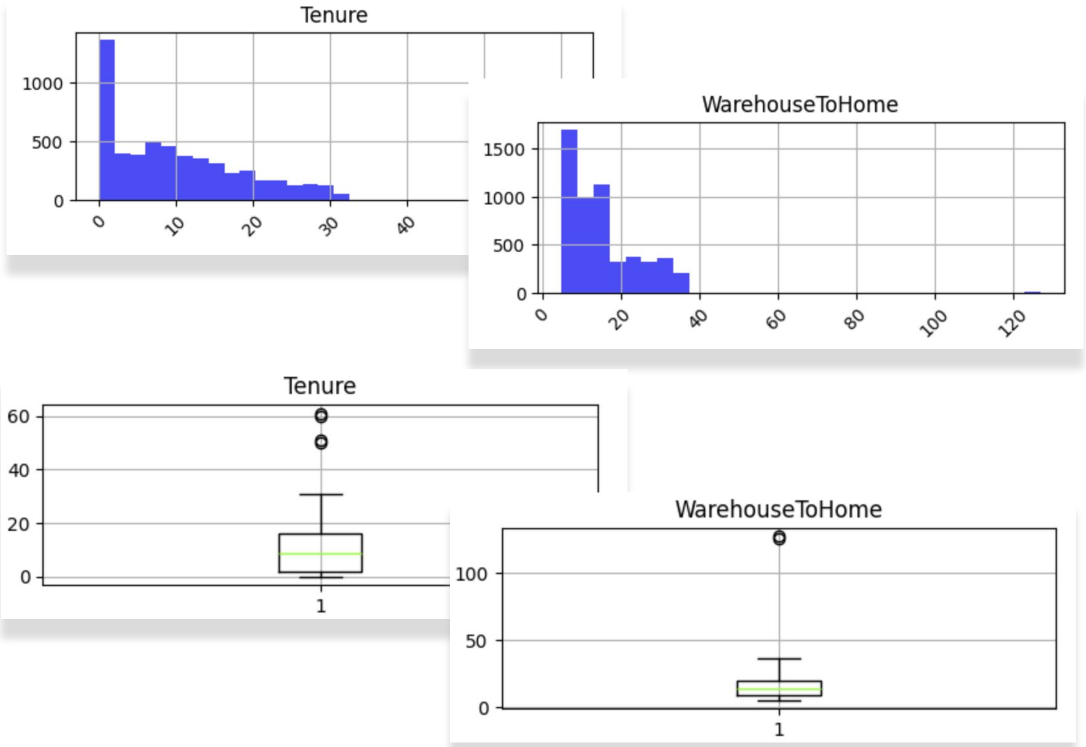
  – *Tenure, WarehouseToHome, NumberOfAddress, etc.*

  Handling: Extracted them with IQRs. I didn’t remove them from analysis to study their impact. *Log-transform, imputation*

- **Categorical Features:**

  – Gender, PreferredLoginDevice, PreferredPaymentMode etc.

  Handling: One-Hot-Encoding for non ordinal data like Gender, Marital Status



Data distribution analysis - Preprocessing

- **Data Variability:**

  – Some features have high variability

  Tenure, WarehouseToHome

  Handling: Used standard-scaler, Log-transform

- **Missing values:**

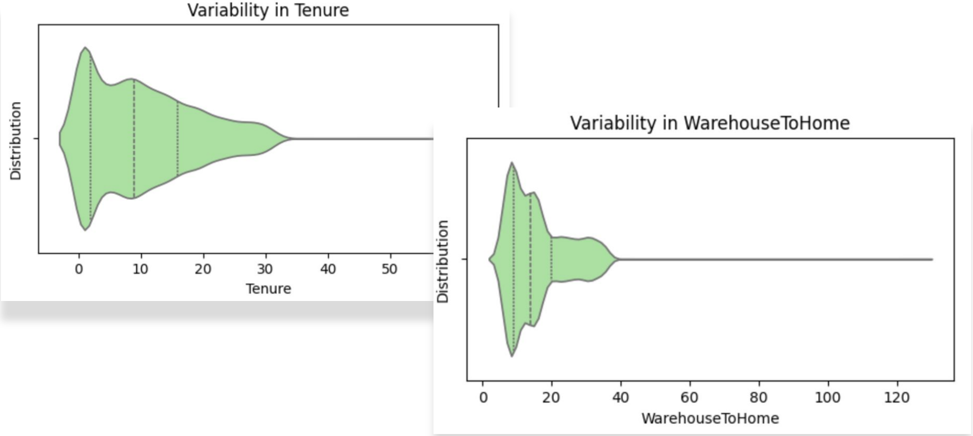
  – *Tenure, WarehouseToHome, HourSpendOnApp, etc.*

  Handling: Used K-Means based imputation.

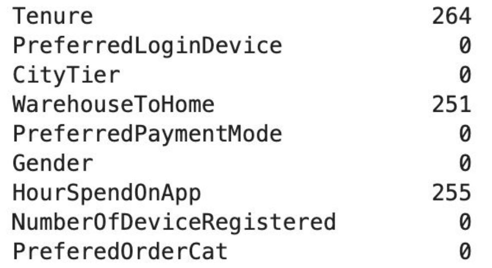
- **Data Imbalance:**

  – Around 5:1 ratio between positive and negative target variable

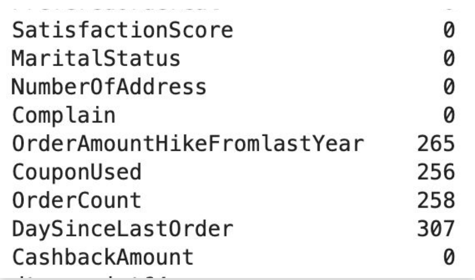
  Handling: SMOTE oversampling



Missing values



```
Tenure                         264
PreferredLoginDevice             0
CityTier                         0
WarehouseToHome                251
PreferredPaymentMode             0
Gender                           0
HourSpendOnApp                 255
NumberOfDeviceRegistered         0
PreferedOrderCat                 0
```



```
SatisfactionScore                0
MaritalStatus                    0
NumberOfAddress                  0
Complain                         0
OrderAmountHikeFromlastYear    265
CouponUsed                     256
OrderCount                     258
DaySinceLastOrder              307
CashbackAmount                   0
```

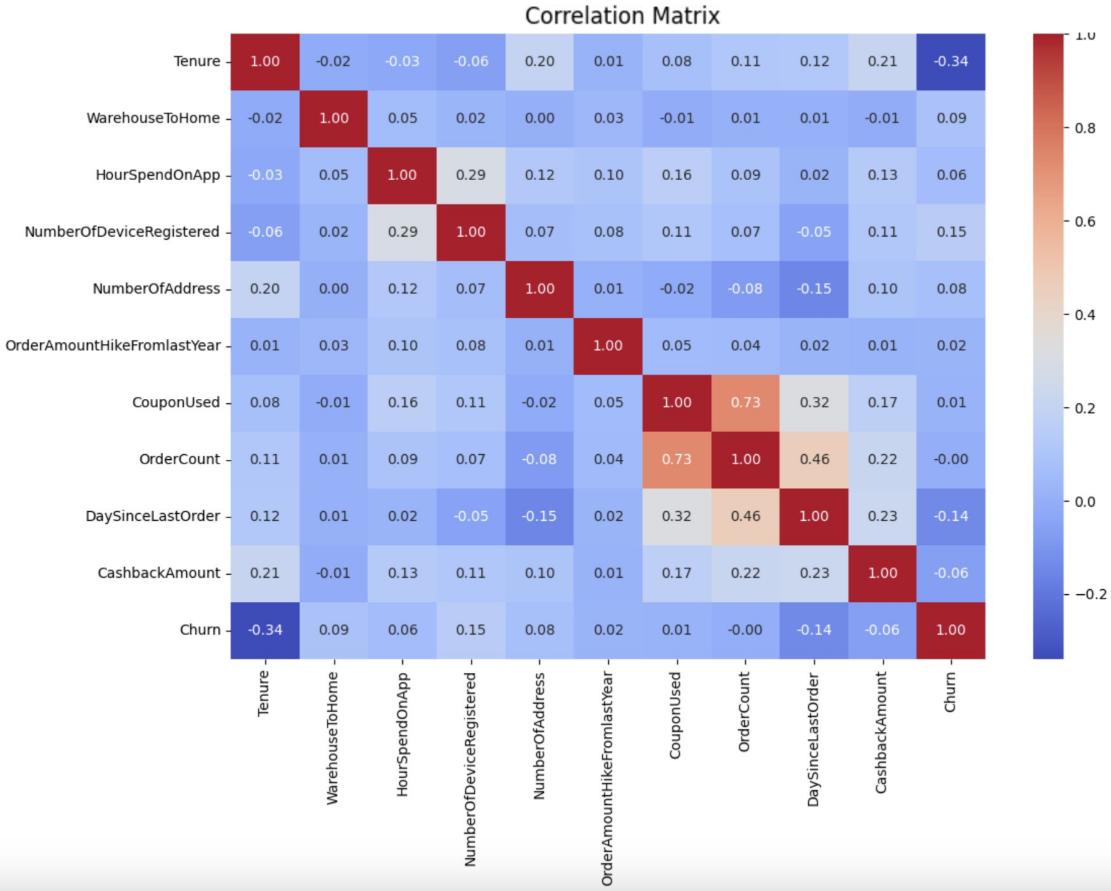
Correlation analysis - to check linear relationship

- ***OrderCount*** and ***CouponUsed*** seem to be highly correlated.

- ***DaySinceLastOrder*** and ***OrderCount*** seem to be moderately correlated.

- *Tenure* seems to be moderately correlated with the target variable.

- Handling:
  - Considering only one of the highly correlated features may help the model.



Correlation Matrix

	Tenure	WarehouseToHome	HourSpendOnApp	NumberOfDeviceRegistered	NumberOfAddress	OrderAmountHikeFromlastYear	CouponUsed	OrderCount	DaySinceLastOrder	CashbackAmount	Churn
Tenure	1.00	-0.02	-0.03	-0.06	0.20	0.01	0.08	0.11	0.12	0.21	-0.34
WarehouseToHome	-0.02	1.00	0.05	0.02	0.00	0.03	-0.01	0.01	0.01	-0.01	0.09
HourSpendOnApp	-0.03	0.05	1.00	0.29	0.12	0.10	0.16	0.09	0.02	0.13	0.06
NumberOfDeviceRegistered	-0.06	0.02	0.29	1.00	0.07	0.08	0.11	0.07	-0.05	0.11	0.15
NumberOfAddress	0.20	0.00	0.12	0.07	1.00	0.01	-0.02	-0.08	-0.15	0.10	0.08
OrderAmountHikeFromlastYear	0.01	0.03	0.10	0.08	0.01	1.00	0.05	0.04	0.02	0.01	0.02
CouponUsed	0.08	-0.01	0.16	0.11	-0.02	0.05	1.00	0.73	0.32	0.17	0.01
OrderCount	0.11	0.01	0.09	0.07	-0.08	0.04	0.73	1.00	0.46	0.22	-0.00
DaySinceLastOrder	0.12	0.01	0.02	-0.05	-0.15	0.02	0.32	0.46	1.00	0.23	-0.14
CashbackAmount	0.21	-0.01	0.13	0.11	0.10	0.01	0.17	0.22	0.23	1.00	-0.06
Churn	-0.34	0.09	0.06	0.15	0.08	0.02	0.01	-0.00	-0.14	-0.06	1.00

Insights and Recommendations

for CX department

(Feature analysis and Engineering)

Understanding The Business

- It seems to be a Walmart/Amazon kind of retail business

- Sells items ranging from ***Electronics*** to ***Fashion*** to ***Grocery***

- Sells items in different city tiers

Customer Journey stages for CX - feature mapping

Retention

- Preferred Payment Mode
- Preferred order category
- Satisfaction scores
- Coupons used
- Complain
- Order Amount Hike
- Marital Status

Engagement

- Hour Spend on App
- Number of Addresses
- Day since last order
- Cashback Amount

Onboarding

- Tenure
- Number of devices registered
- Warehouse to Home

Acquisition

- Preferred Login Device
- City Tier
- Gender

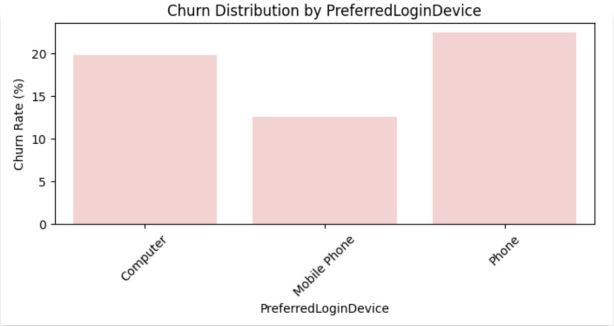
By categorizing these features based on the customer journey stages, the CX department can:

- Better understand how each feature contributes to the overall customer experience
- Identify opportunities for improvement and retention.

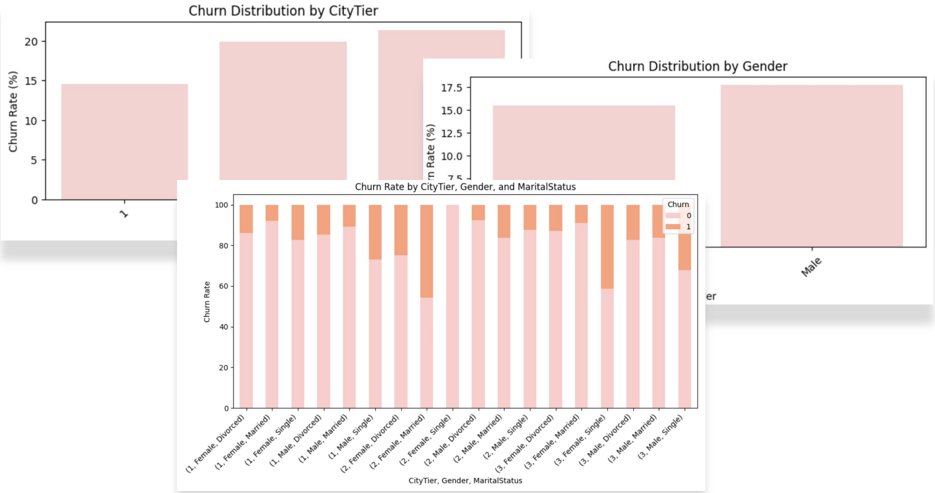
Acquisition

(Preferred Login device, Demography)

- Preferred Login Device:
  - **Insights:** Customers with 'Phone' and 'Computer' as Preferred Login Device have higher churn rates.
  - **Recommendation:** Review the computer and phone experiences and optimize as needed.



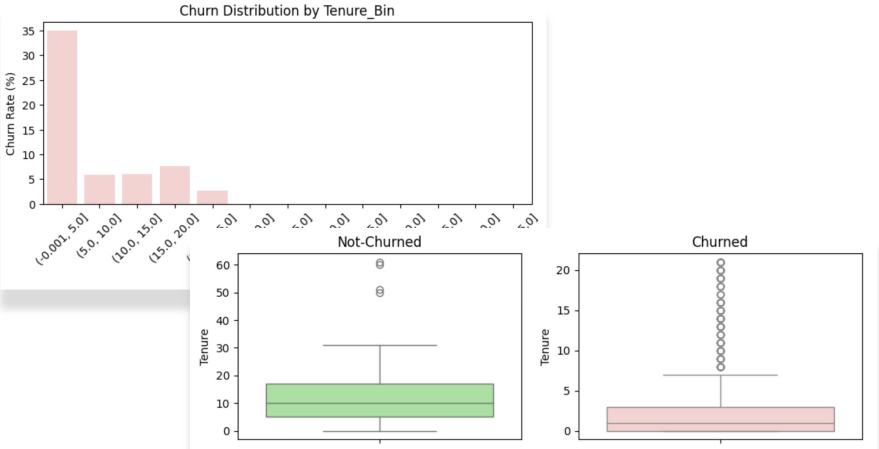
- Demographic factors (city tier, gender and marital status):
  - **Insights:**
    - Tier 2 and 3 cities seem to be more affected by churn
    - Gender alone seems to be neutral
    - Females from city tier-2 are most affected
    - Analyzing churn rate for all demographic factors together, following groups seem to be affected the most:
      - Married females from city tier - 2 and Single females from city tier - 3
      - Single males from city tier - 1 and 3
      - Divorced females from city tier - 2
  - **Recommendation:**
    - Tailoring marketing strategies and service offerings around a combination of these insights may help improve customer stickiness.



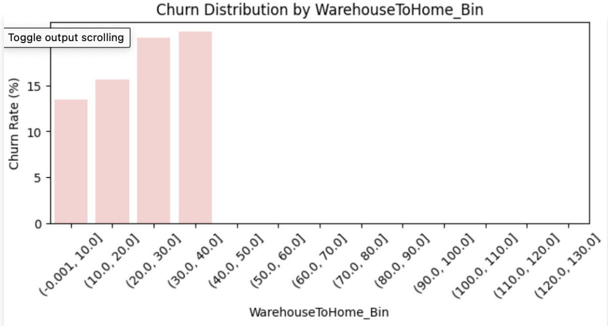
Onboarding

(Tenure, Number of devices, Warehouse to home)

- Tenure:
  - **Insights:**
    - Customers tend to churn early in their tenure than later
    - The median tenure for churned customers is around 1 unit, for not-churned customers it is 10 units of time.
  - **Recommendation:** Review the customer onboarding process or early engagement strategies.



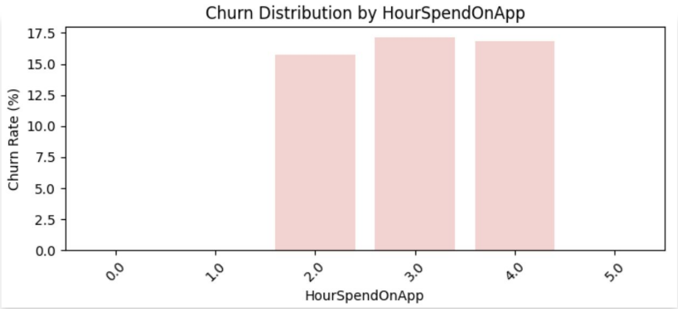
- Warehouse to home:
  - **Insights:**
    - Churn rates seem to go higher as the distance increases.
  - **Recommendation:**
    - Logistics involved in delivering the product etc. may need review which might be affecting the customer onboarding experience.



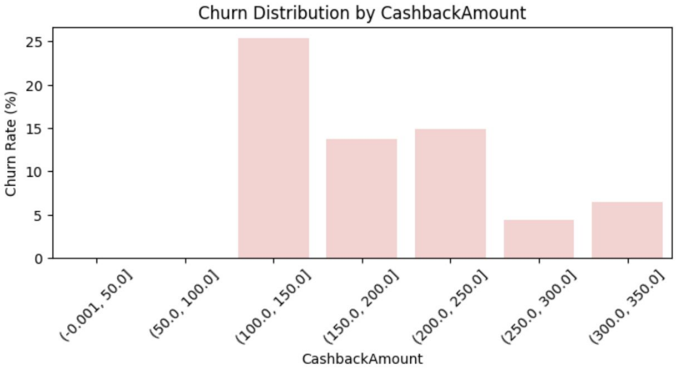
Engagement

(Hour Spend on App, Number of Addresses, Day since last order, Cashback Amount)

- Hours Spent on App:
  - **Insights:**
    - Customers seem to spend a good amount of time on the app.
    - Churn is pretty consistent across different hours of app usage, so no concrete insights can be drawn out of it.
  - **Recommendation:** All good when viewed stand alone.



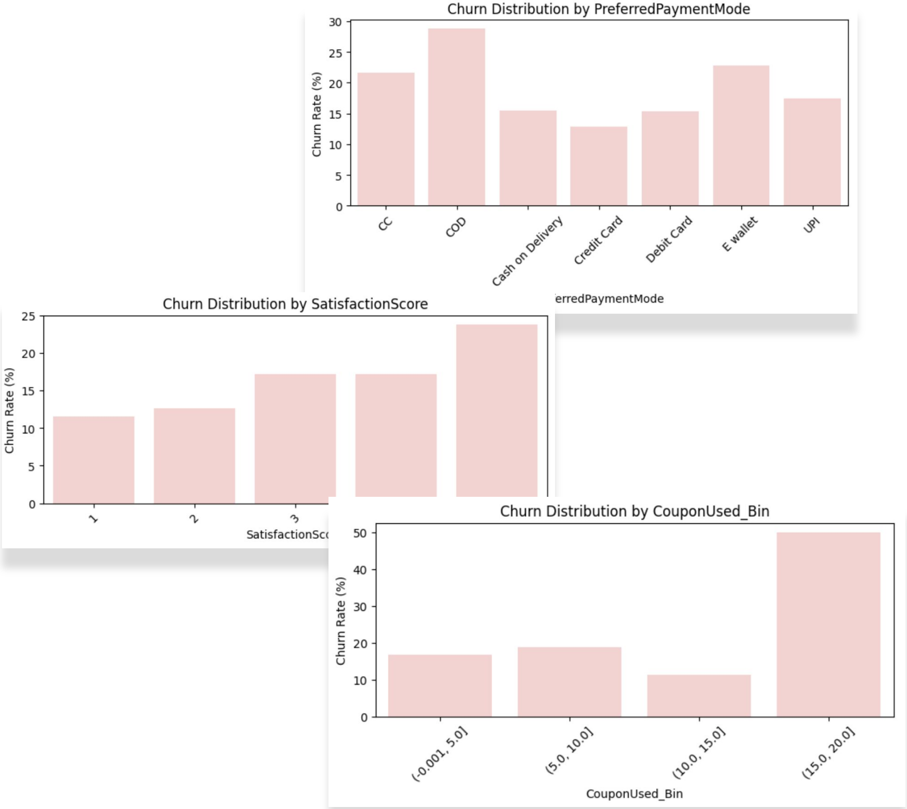
- Cashback Amount:
  - **Insights:**
    - Churn rates seem to be lower as the amount of cashback increases.
    - Reflects the incentives or rewards offered to customers, which can influence their repeat purchases and loyalty.
  - **Recommendation:**
    - Higher cashback seems to be working to keep the customers, so a personalized cashback or other reward programme should be explored for the churning segment.



Retention

(Preferred Payment Mode, Satisfaction scores, Coupons used etc.)

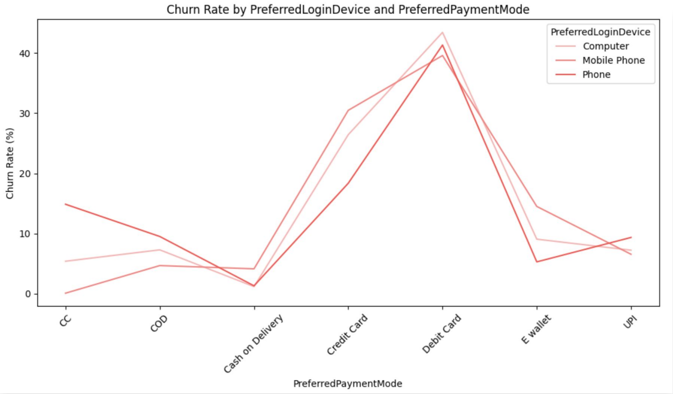
- Preferred Payment Mode:
  - **Insights:**
    - Churn rate around COD, CC and E-wallet are higher than other modes.
  - **Recommendation:**
    - CX might want to encourage customers towards signing up for credit card, debit card etc. payment modes for better customer stickiness.
- Satisfaction scores:
  - **Insights:**
    - Assuming 5 being the worst satisfaction score, the churn rate is pretty aligned with the satisfaction score.
  - **Recommendation:**
    - Further qualitative analysis of the feedback data (text analysis with NLP, GenAI/LLMs etc.) may highlight subtleties of customer experience and potential improvement areas.
- Coupons used:
  - **Insights:**
    - The order count, coupons used and order amount hike from last year all together give the same insights, customers who have higher of these features also have higher churn rates
  - **Recommendation:**
    - It might be useful to review pricing and promotions. Strategies should be thought of around how to motivate these customers to continue with us.



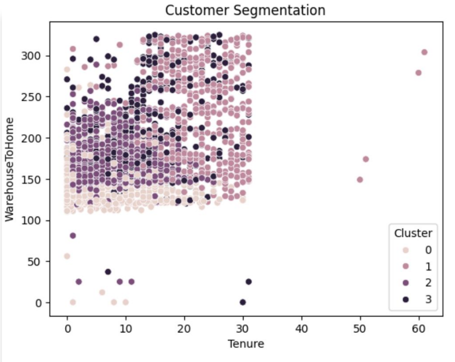
Cross-category interactions between features

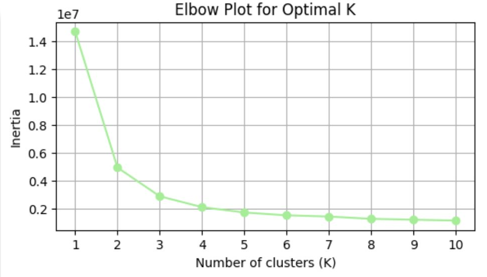
Feature Generation

- Preferred Login Device and Preferred Payment Mode:
  - **Insights:**
    - Customers with preferred_payment_mode as ‘Credit card’ and preferred_login_device as 'Mobile Phone' has higher churn than with Computer or Phone.
  - **Recommendation:**
    - CX may want to review "credit card" payment process on Mobile phone.



- Customer Segmentation using K-Nearest Neighbors clustering
  - Created ***customer segment clusters*** based on some original features and used that as a new feature for training prediction models.

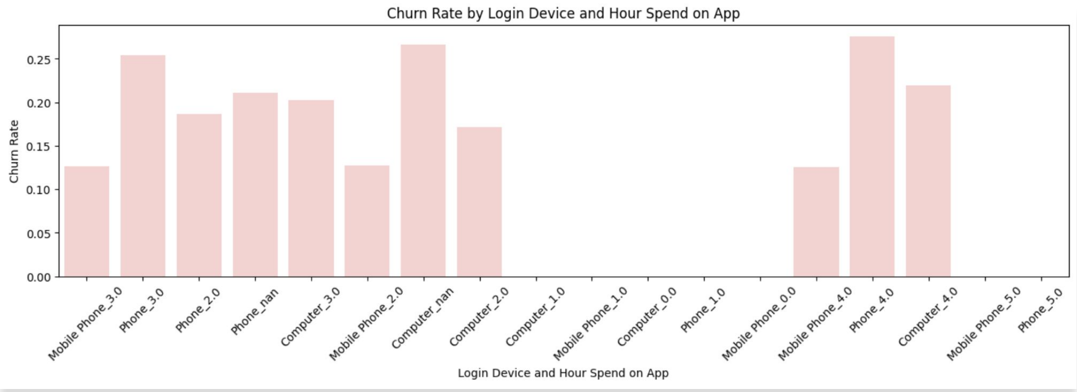


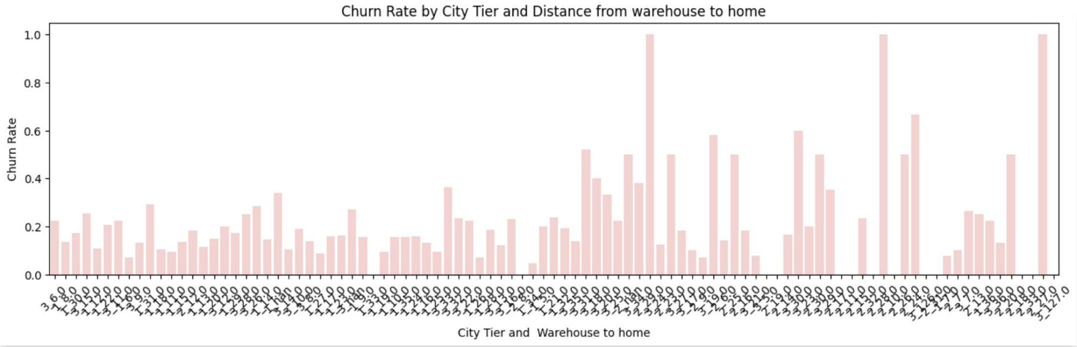


Cross-category interactions between features

Feature Generation

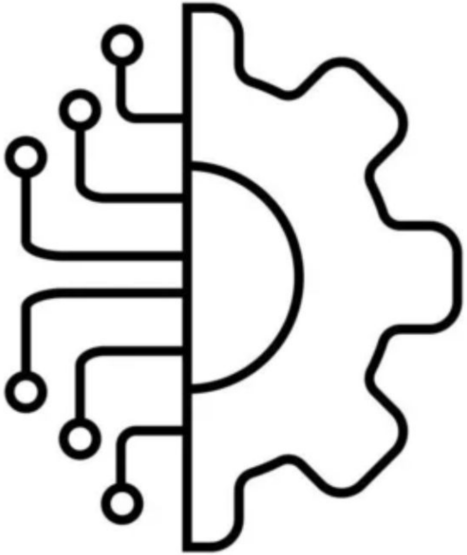
- Sum/Difference interactions:
  - Preferred Login Device ***and*** Hour Spend On App
  - Preferred Payment Mode ***and*** Order Amount Hike From last Year
  - City Tier ***and*** Warehouse To Home
  - Number Of Devices Registered ***and*** Order Count
- Ratio interactions:
  - Hour Spend On App ***and*** Order Count
  - Number Of Devices Registered ***and*** Days since Last Order
- Product or Polynomial interactions:
  - Satisfaction Score ***and*** Order Amount Hike From last Year
  - Number Of Address ***and*** Number Of Device Registered





Churn Prediction Modeling

Machine Learning



Machine Learning - Algorithms, Metrics and Training

- ML based binary classification modeling was applied to predict churn:
    - Used Binary Classification Algorithms used:
        - Logistic regression: as a base linear model
        - Support Vector Machines (SVM)
        - Tree based models:
            - Random Forest
            - XGBoost
        - Simple fully connected neural-net and Multi-level Perceptron Classifier (MLP from scikit library)
    - Performance metrics:
        - Precision, Recall, F1, ROC_AUC, micro and macro F1
        - Since churn prediction has data-imbalance, this set of metrics are good at evaluating such scenarios. F1 and ROC_AUC provide a balanced approach to evaluating minority and majority classes
    - Training Approach:
        - Cross validation for training and evaluation
        - Grid Search for hyper-parameter tuning

- Best performing model came out to be: XGBoost (with a ROC_AUC score of 0.95 on test data)
  [Precision: 1, Recall: 0.90, F1: 0.94]
- Neural-net and Random Forest were close too
- **Note:** I couldn’t test much of features engineered due to lack of time. The models can be further enhanced.

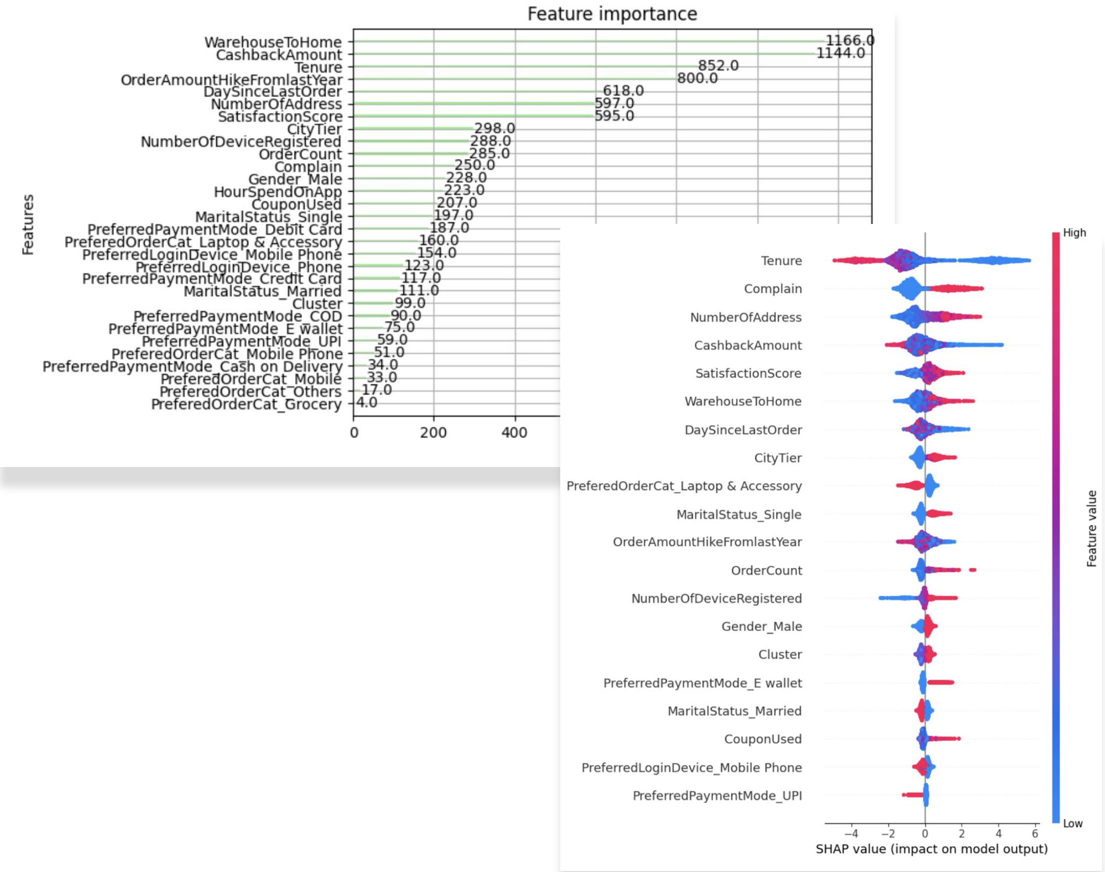
Feature Importance and feature selection

- The left plot shows the feature importance for the best performing XGBoost model

  – Fscore represents the number of times a feature appears in a tree across all boosting rounds

- The ***SHAP values*** in the other plot indicate the feature importance and how they impact churn

- The model/s can be further simplified/enhanced by retraining with only the selected important features



Next Steps

- Further study of pairs of features to understand their relationship.

- Features developed - **standalone and interacting** - can be used to train models and further enhance their performance.

- Data imbalance handling can also be explored further using techniques like SMOTE.

- Modeling **Time-to-churn** (Survival Analysis)
    - COX-Proportional Hazard model

Thank you!