Churn Prediction

Data-Driven Insights for GTM, and Product

Objective

- Analyze the provided dataset: Customer Churn
- Develop actionable insights and recommendations for CX

Data Exploration

- Load and preprocess the dataset
- Perform exploratory data analysis (EDA) for the dataset
- Visualize key features and relationships

Features:

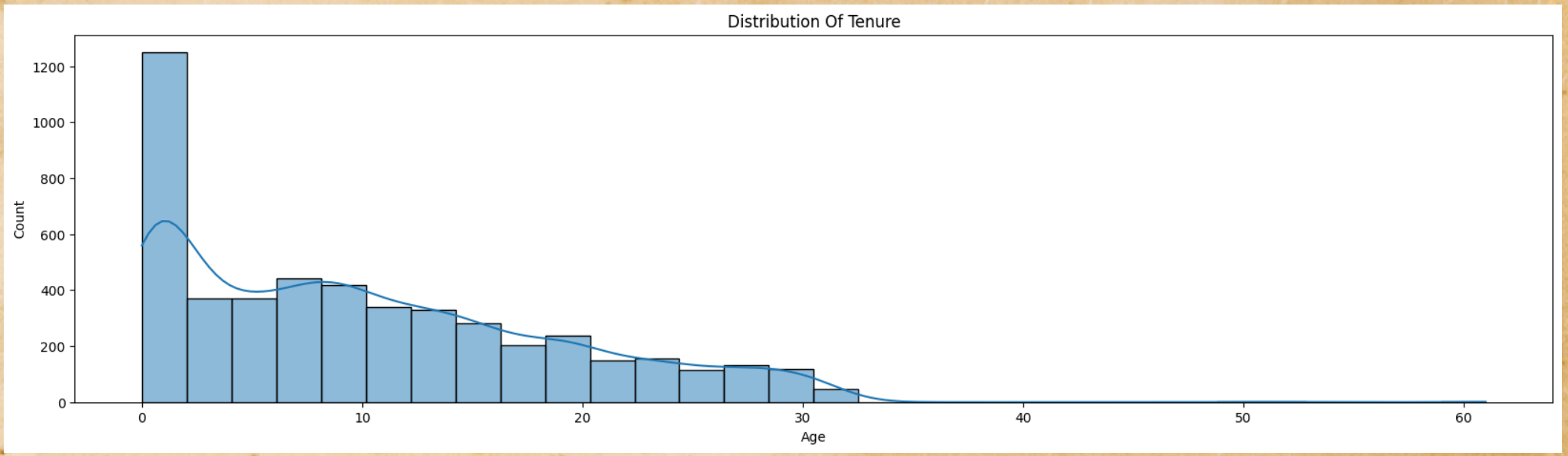
- Categorical:
  - OrderCat, Gender, PreferredLoginDevice, MaritalStatus, PaymentMode
- Numeric:
  - Churn, Tenure, CityTier, WarehouseToHome, HourSpendOnApp, NumberOfDeviceRegistered, SatisfactionScore, NumberOfAddress, Complain, OrderAmountHikeFromlastQuarter, CouponUsed, OrderCount, DaySinceLastOrder, CashbackAmount

EDA - 1.1 Tenure Distribution vs Churn Rate Analysis

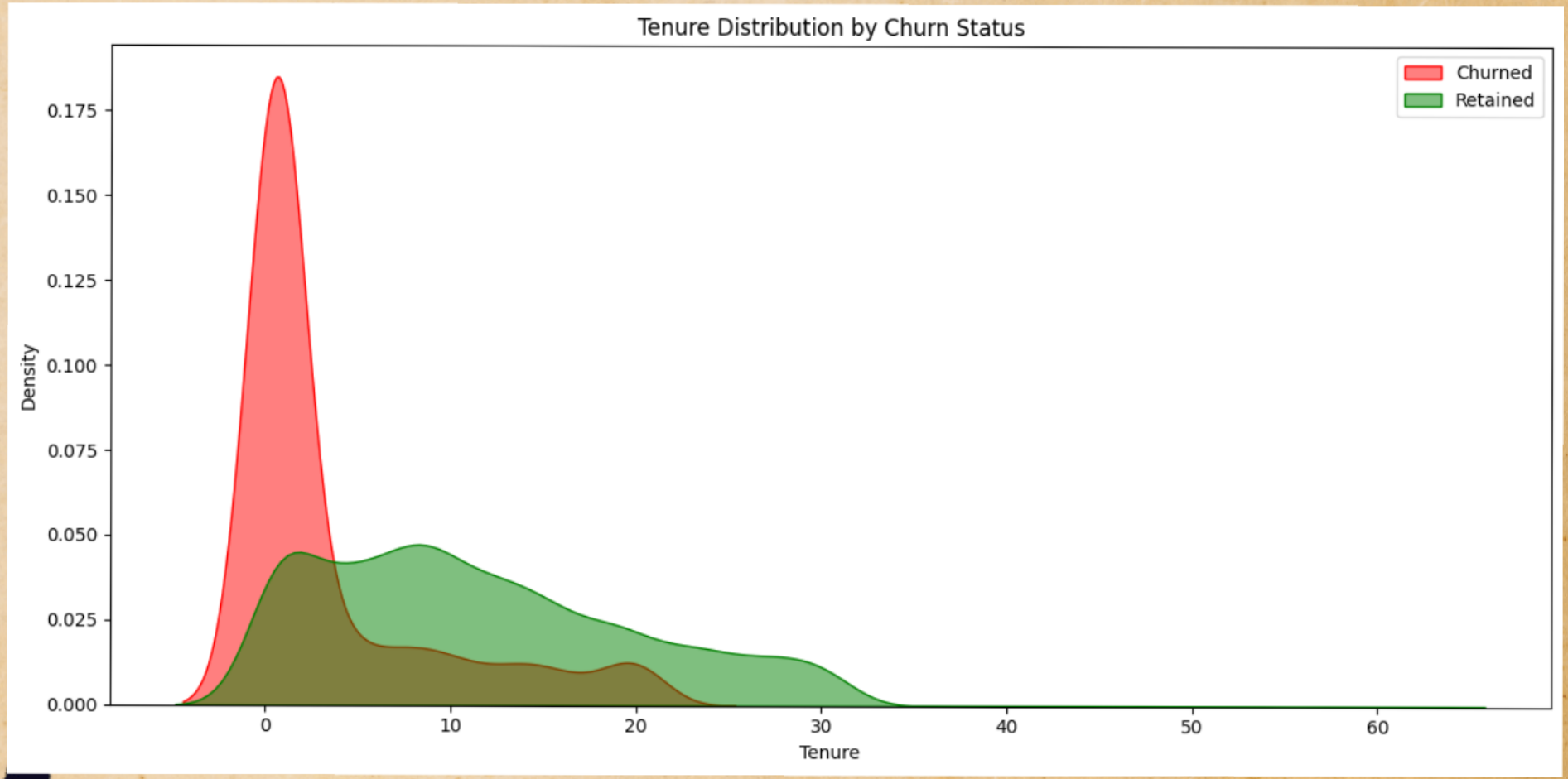
Tenure Statistics

- count 4964.000000
- mean 10.276793
- std 8.630337
- min 0.000000
- 25% 2.000000
- 50% 9.000000
- 75% 16.000000
- max 61.000000
- .
- .

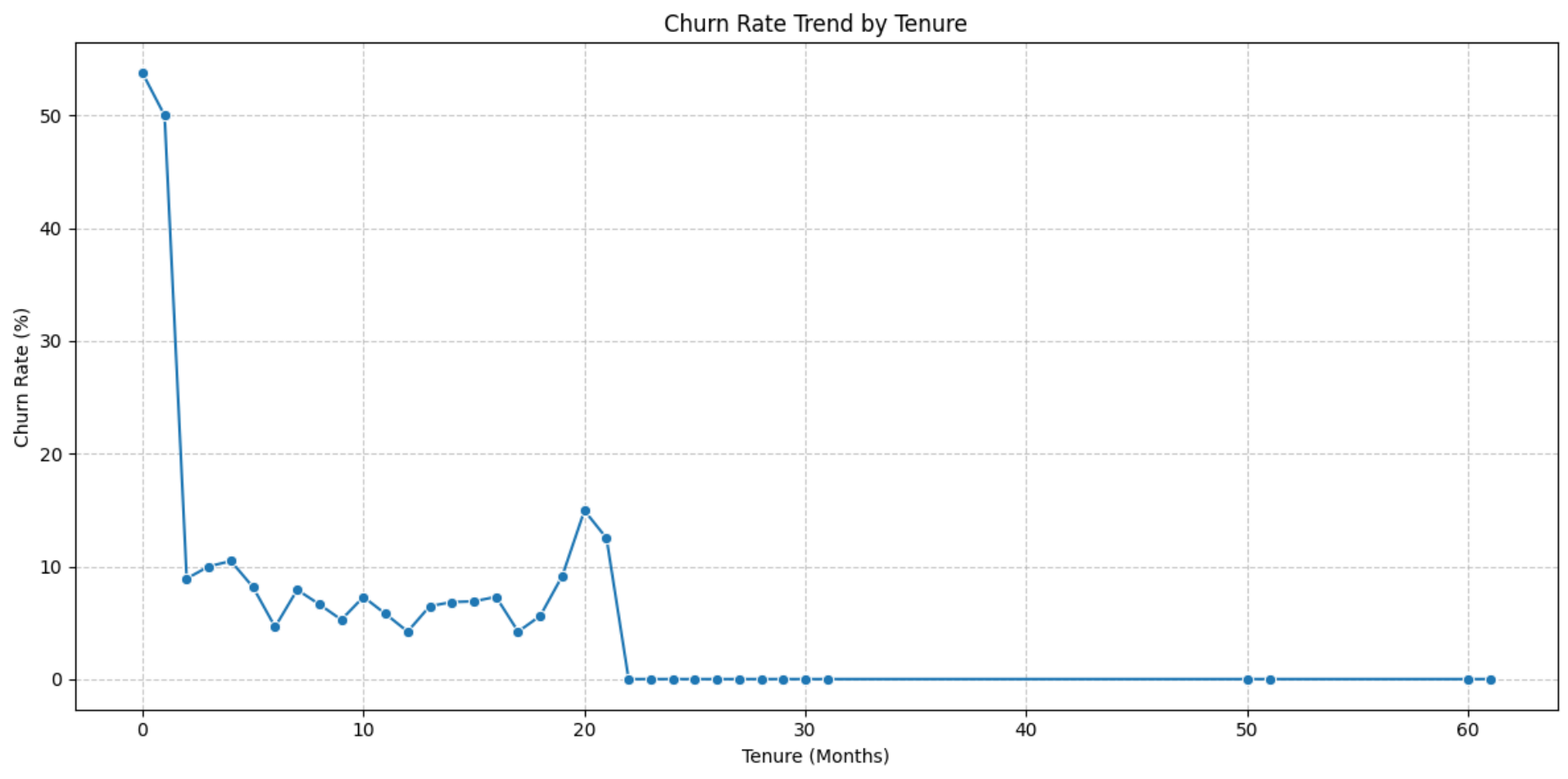
Distribution of Tenure



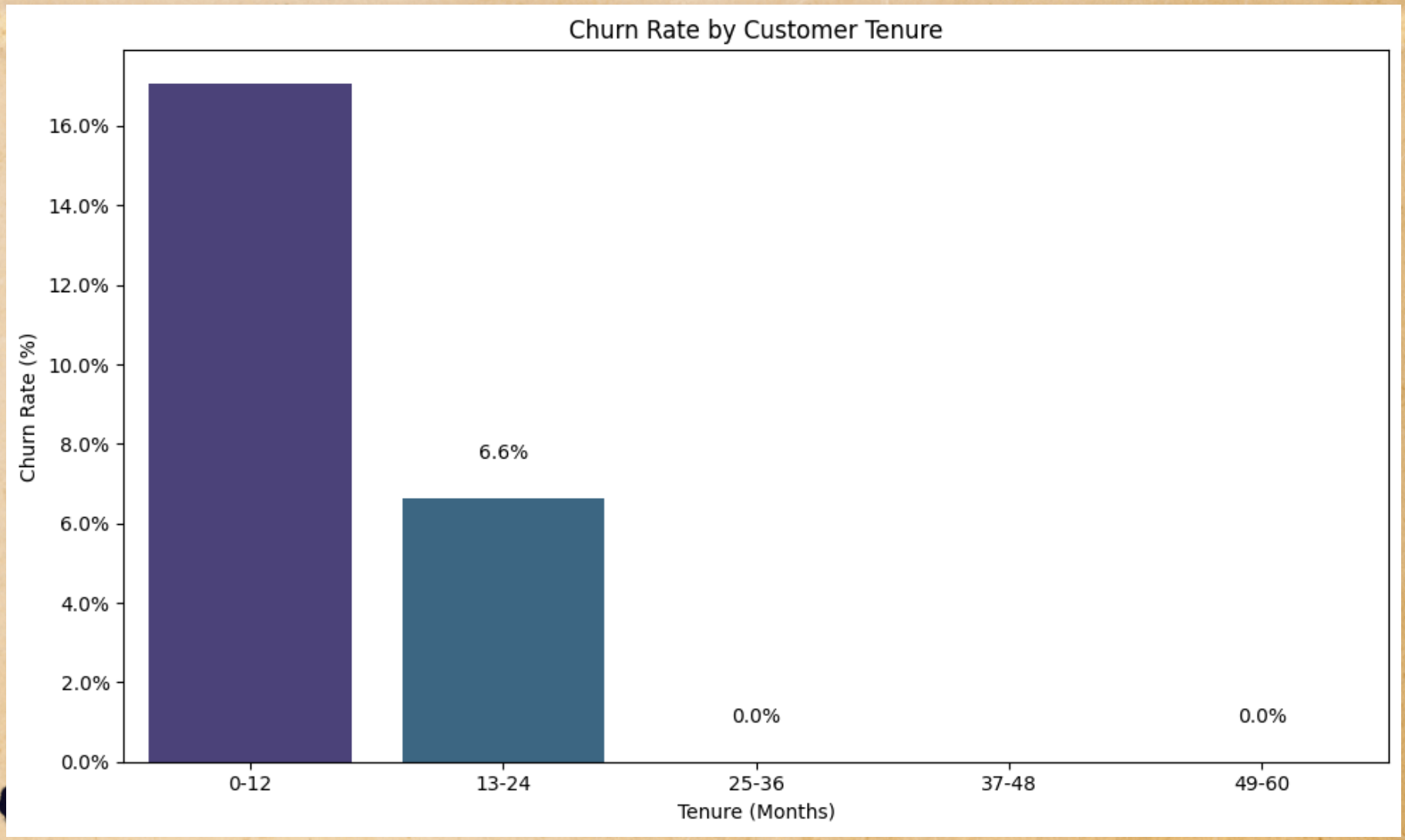
Kernel Density Estimate



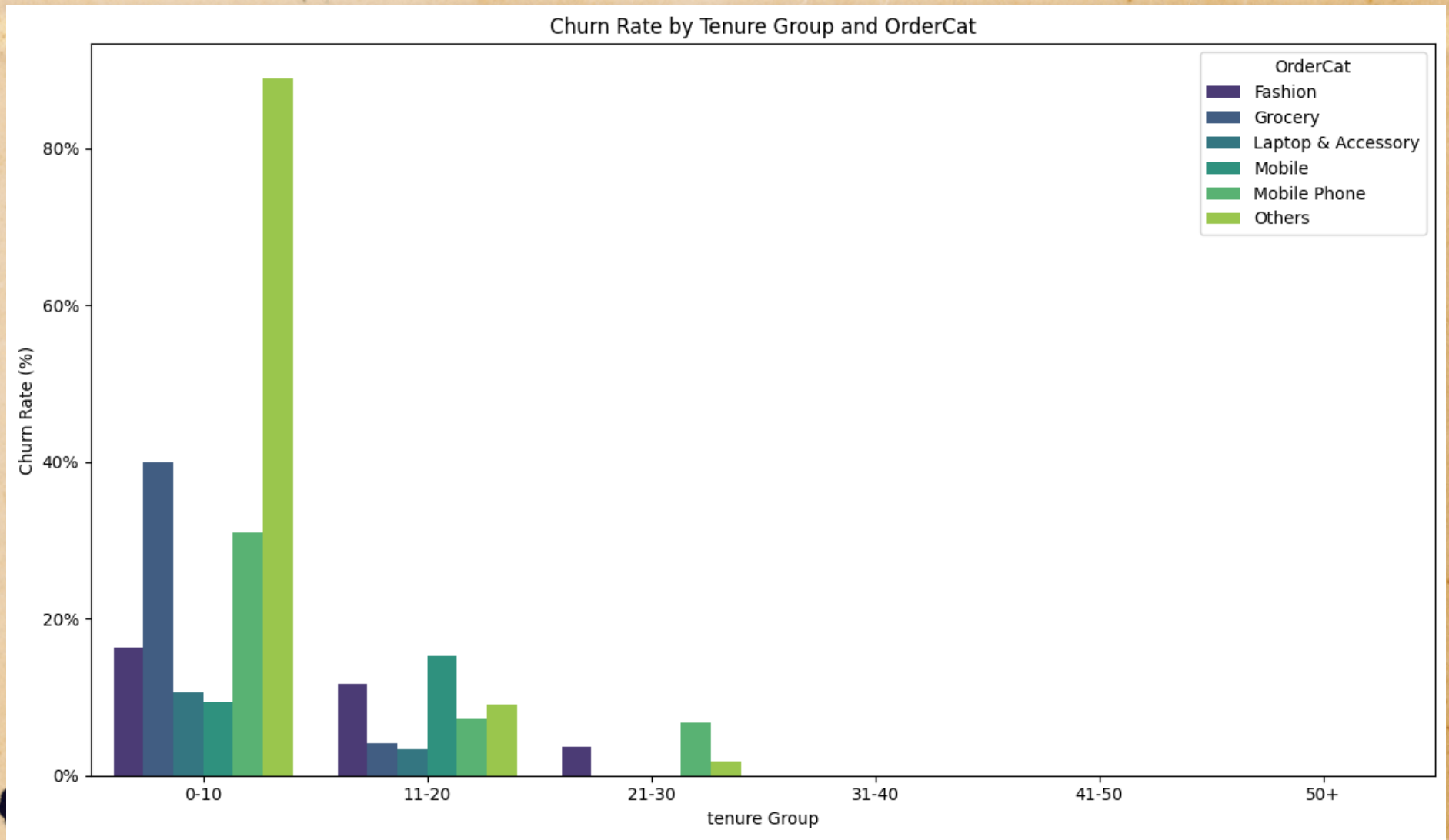
Churn rate for each tenure month

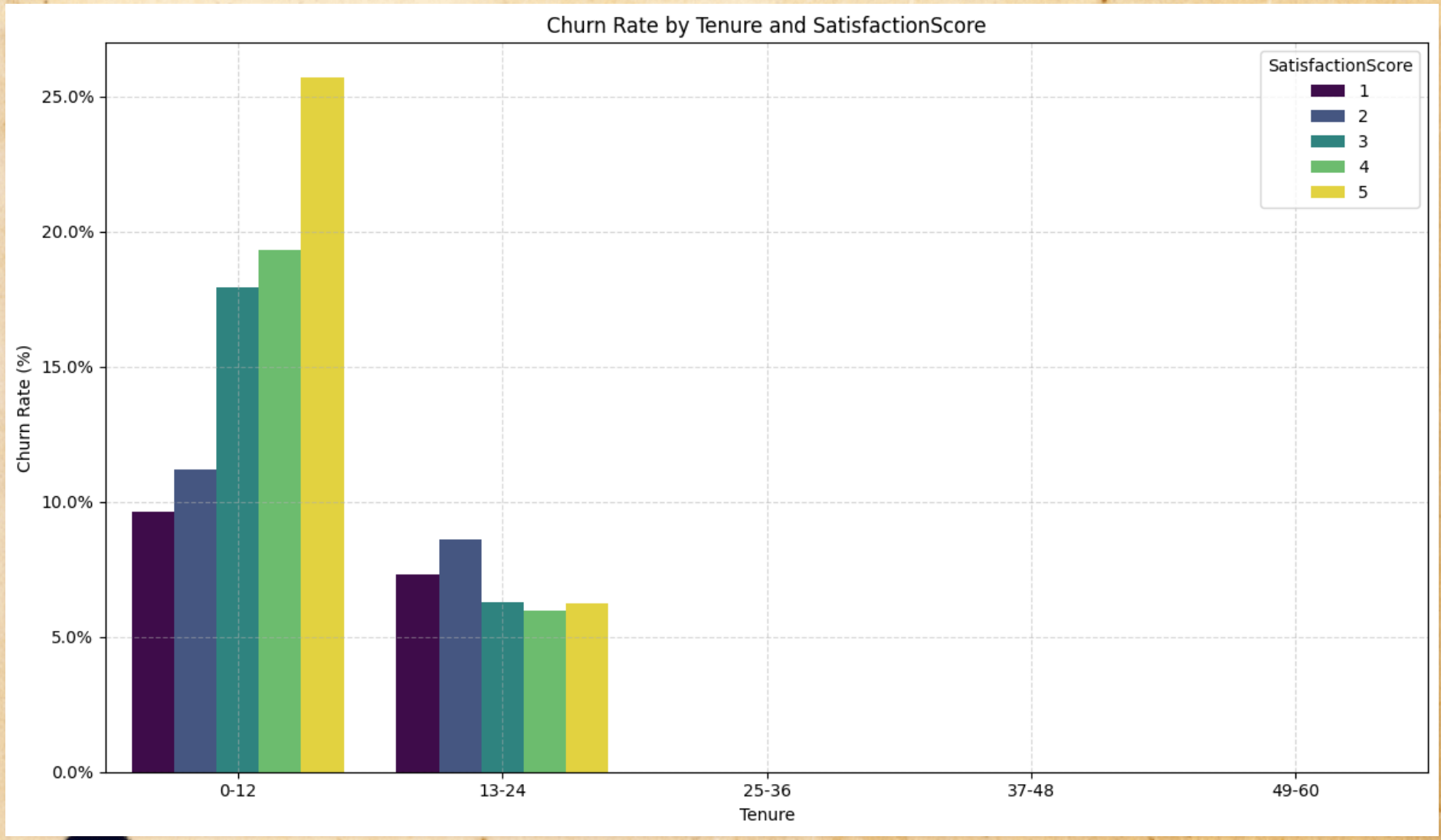


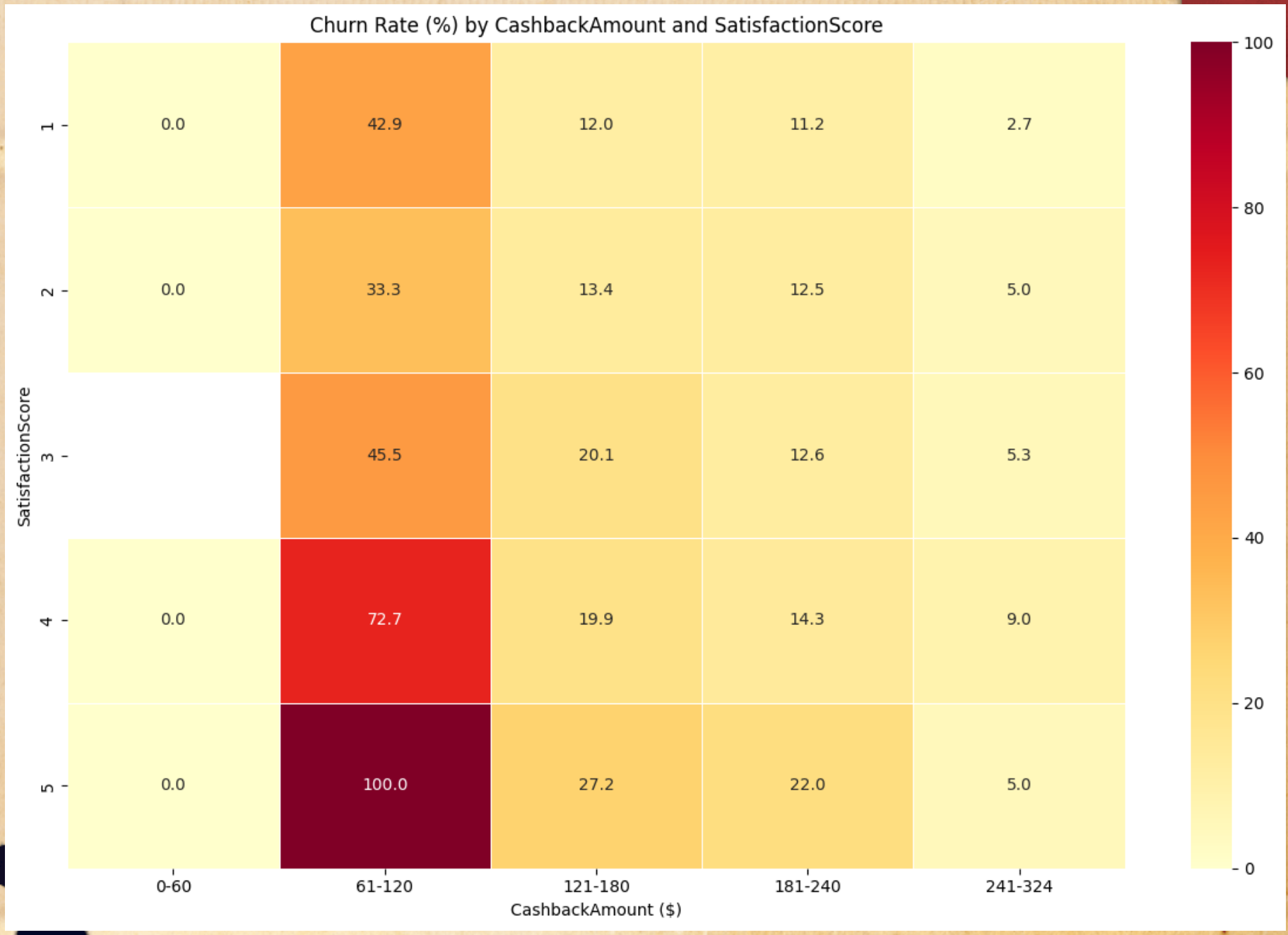
Tenure Group: 0-12, 13-24, 25-36, 37-48, 49-60



Churn Rate by multiple features

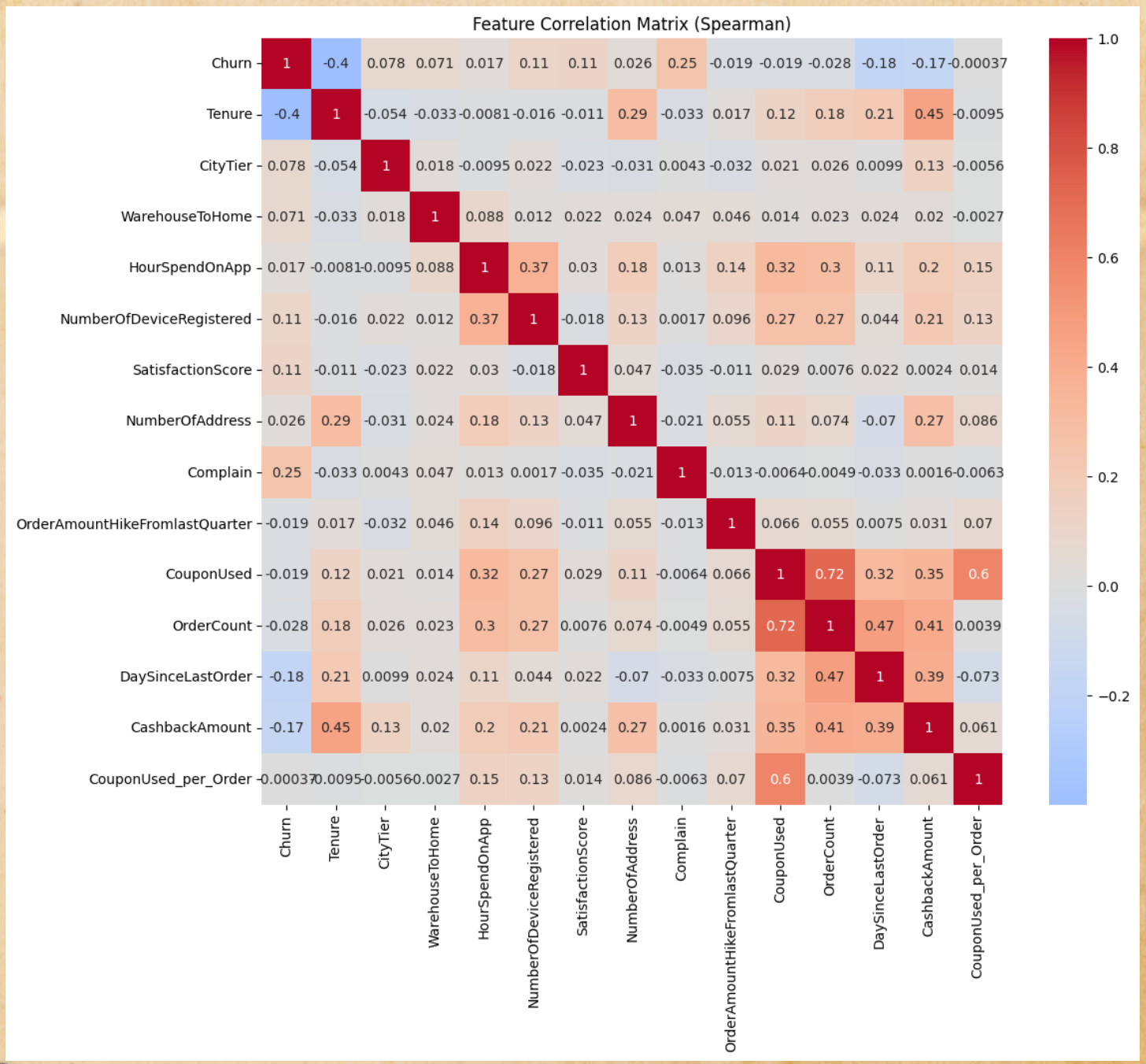



Churn Rate by Tenure and SatisfactionScore
SatisfactionScore
1
2
3
4
5
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
Churn Rate (%)
0-12
13-24
25-36
37-48
49-60
Tenure


Churn Rate (%) by CashbackAmount and SatisfactionScore
SatisfactionScore
1
2
3
4
5
0.0
42.9
12.0
11.2
2.7
0.0
33.3
13.4
12.5
5.0
45.5
20.1
12.6
5.3
0.0
72.7
19.9
14.3
9.0
0.0
100.0
27.2
22.0
5.0
0-60
61-120
121-180
181-240
241-324
CashbackAmount ($)
100
80
60
40
20
0

EDA Statistical Correlation Analysis

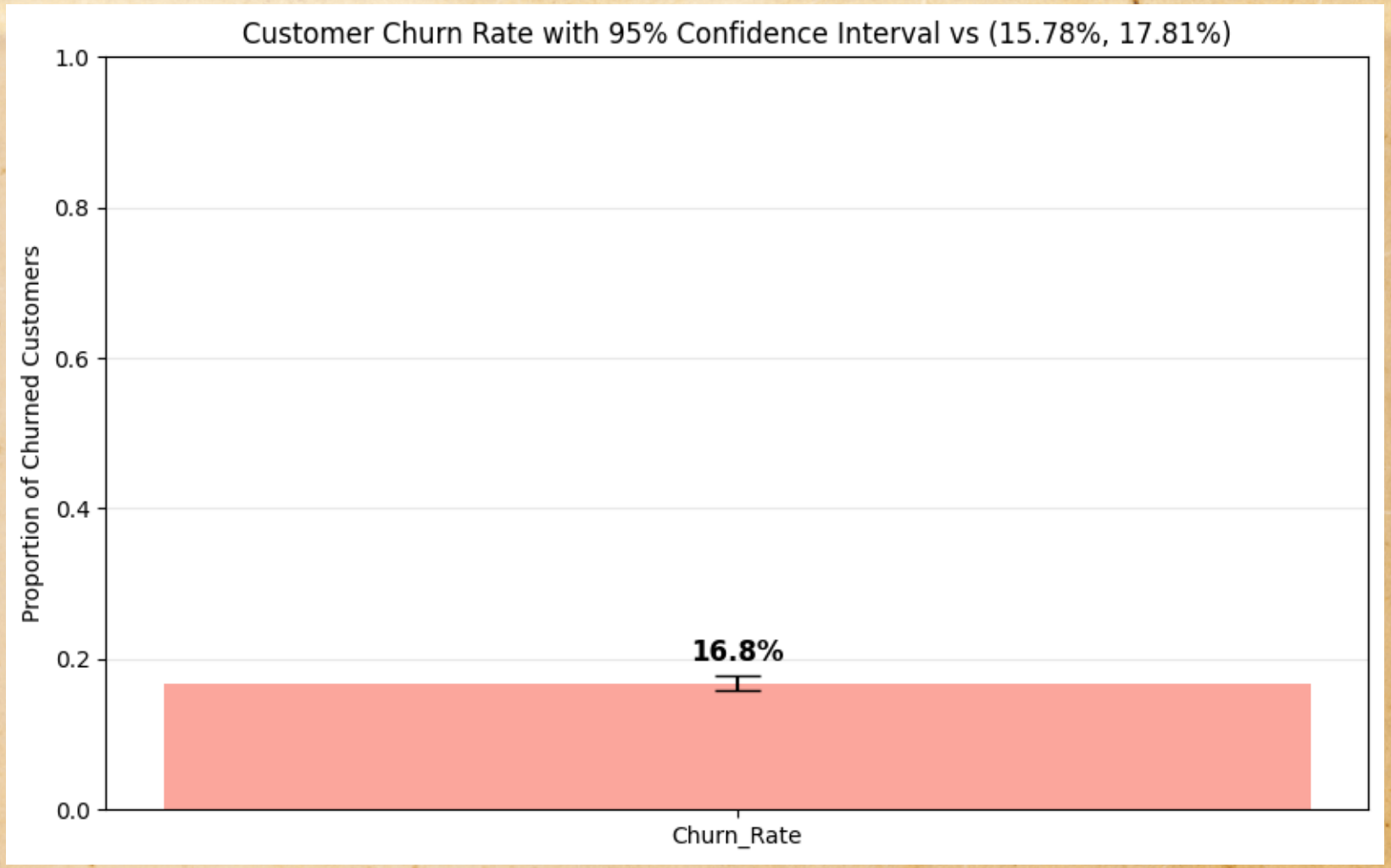
- Numeric Features

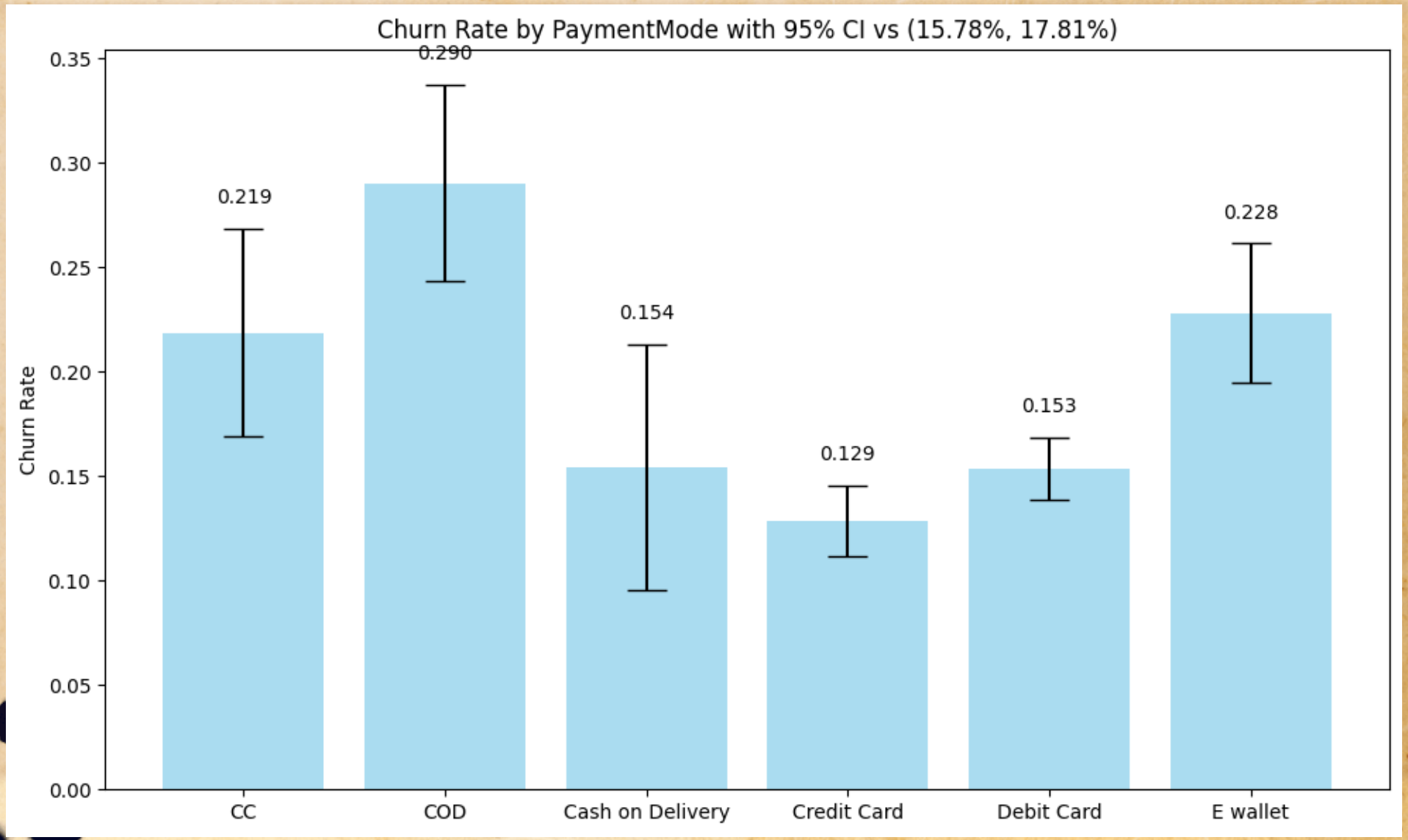

Feature Correlation Matrix (Spearman)
Churn 1 -0.4 0.078 0.071 0.017 0.11 0.11 0.026 0.25 -0.019 -0.019 -0.028 -0.18 -0.17 -0.00037
Tenure -0.4 1 -0.054 -0.033 -0.0081 -0.016 -0.011 0.29 -0.033 0.017 0.12 0.18 0.21 0.45 -0.0095
CityTier 0.078 -0.054 1 0.018 -0.0095 0.022 -0.023 -0.031 0.0043 -0.032 0.021 0.026 0.0099 0.13 -0.0056
WarehouseToHome 0.071 -0.033 0.018 1 0.088 0.012 0.022 0.024 0.047 0.046 0.014 0.023 0.024 0.02 -0.0027
HourSpendOnApp 0.017 -0.0081 -0.0095 0.088 1 0.37 0.03 0.18 0.013 0.14 0.32 0.3 0.11 0.2 0.15
NumberOfDeviceRegistered 0.11 -0.016 0.022 0.012 0.37 1 -0.018 0.13 0.0017 0.096 0.27 0.27 0.044 0.21 0.13
SatisfactionScore 0.11 -0.011 -0.023 0.022 0.03 -0.018 1 0.047 -0.035 -0.011 0.029 0.0076 0.022 0.0024 0.014
NumberOfAddress 0.026 0.29 -0.031 0.024 0.18 0.13 0.047 1 -0.021 0.055 0.11 0.074 -0.07 0.27 0.086
Complain 0.25 -0.033 0.0043 0.047 0.013 0.0017 -0.035 -0.021 1 -0.013 -0.0064 -0.0049 -0.033 0.0016 -0.0063
OrderAmountHikeFromlastQuarter -0.019 0.017 -0.032 0.046 0.14 0.096 -0.011 0.055 -0.013 1 0.066 0.055 0.0075 0.031 0.07
CouponUsed -0.019 0.12 0.021 0.014 0.32 0.27 0.029 0.11 -0.0064 0.066 1 0.72 0.32 0.35 0.6
OrderCount -0.028 0.18 0.026 0.023 0.3 0.27 0.0076 0.074 -0.0049 0.055 0.72 1 0.47 0.41 0.0039
DaySinceLastOrder -0.18 0.21 0.0099 0.024 0.11 0.044 0.022 -0.07 -0.033 0.0075 0.32 0.47 1 0.39 -0.073
CashbackAmount -0.17 0.45 0.13 0.02 0.2 0.21 0.0024 0.27 0.0016 0.031 0.35 0.41 0.39 1 0.061
CouponUsed_per_Order -0.00037 0.0095 -0.0056 -0.0027 0.15 0.13 0.014 0.086 -0.0063 0.07 0.6 0.0039 -0.073 0.061 1
Churn
Tenure
CityTier
WarehouseToHome
HourSpendOnApp
NumberOfDeviceRegistered
SatisfactionScore
NumberOfAddress
Complain
OrderAmountHikeFromlastQuarter
CouponUsed
OrderCount
DaySinceLastOrder
CashbackAmount
CouponUsed_per_Order
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

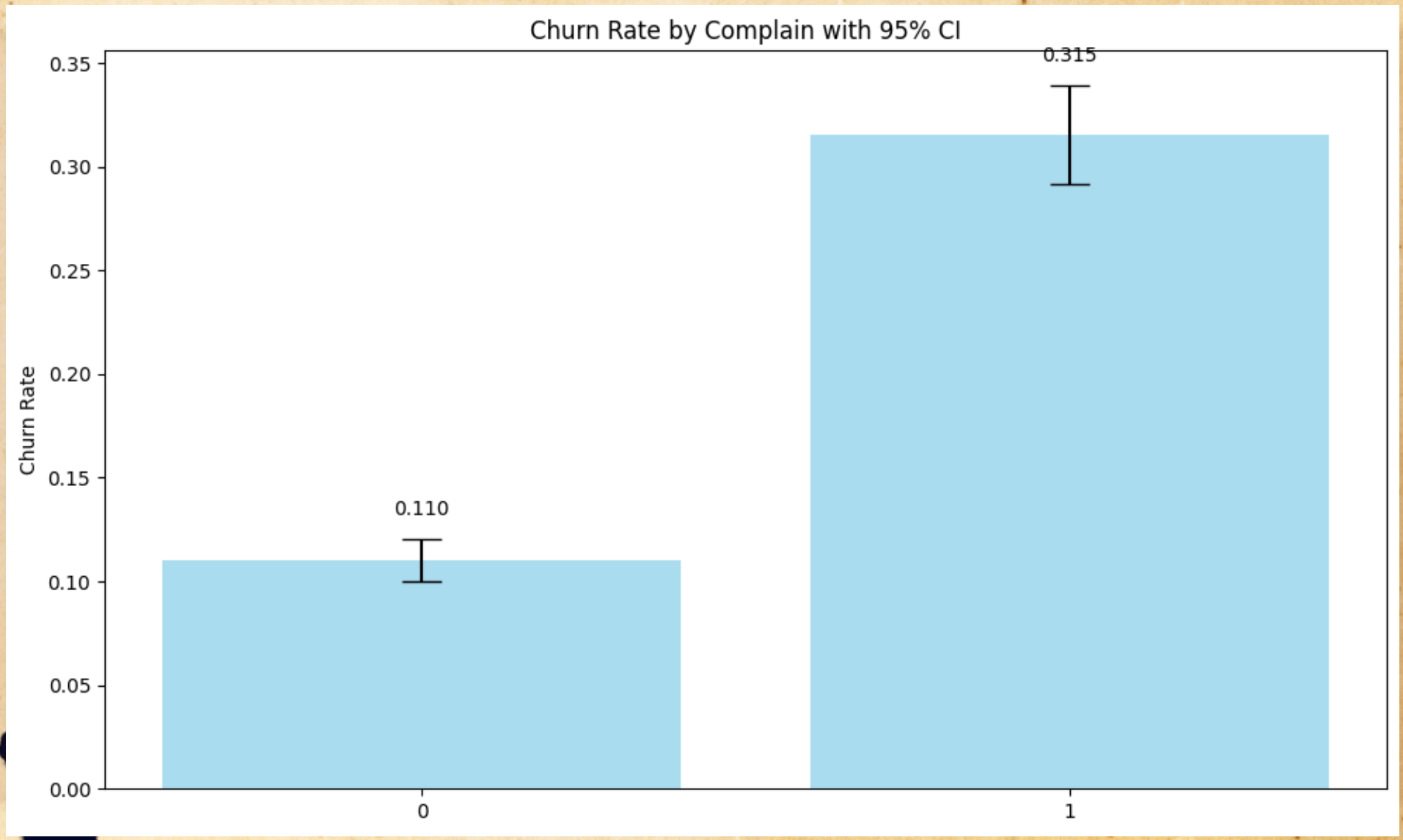
EDA Statistical Correlation Analysis

- Categorical Features

Confidence Interval churn rate



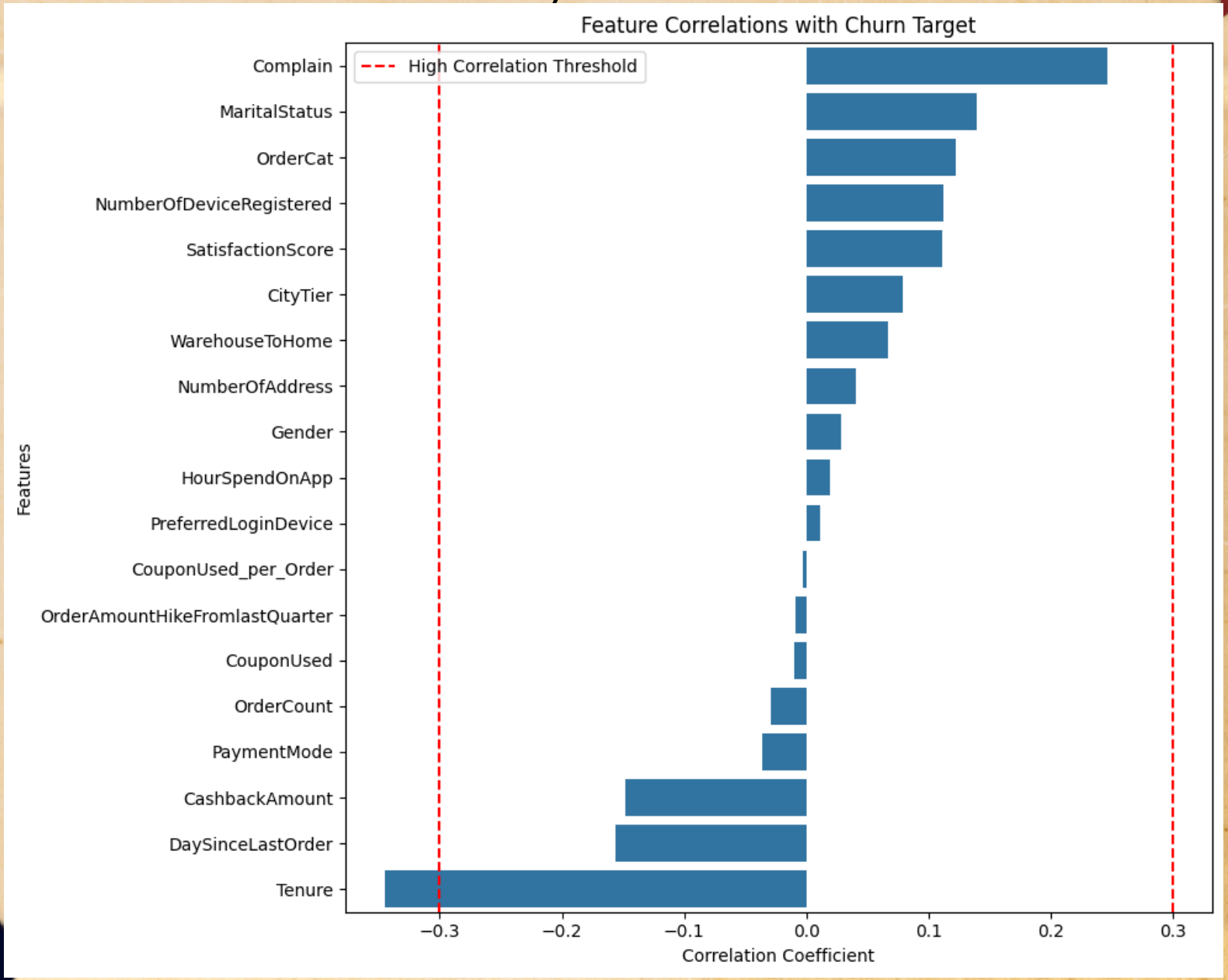

Churn Rate by PaymentMode with 95% CI vs (15.78%, 17.81%)
Churn Rate
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.219
0.290
0.154
0.129
0.153
0.228
CC
COD
Cash on Delivery
Credit Card
Debit Card
E wallet


Churn Rate by Complain with 95% CI
Churn Rate
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.110
0.315
0
1

Statistical-test for Complain Distribution by Churn Status

- T-test Results:
- T-statistic: 16.45
- P-value: 0.0000
-
- Chi-square Results:
- Chi2-stat: 315.88
- P-value: 0.0000

Correlation Analysis



Feature Selection

- Select features with corr>=0.05 and corr<=0.7
  - Complain
  - MaritalStatus
  - OrderCat
  - NumberOfDeviceRegistered
  - SatisfactionScore
  - CityTier
  - WarehouseToHome
  - CashbackAmount
  - DaySinceLastOrder
  - Tenure

Model Selection

- Logistic Regression

  pro: simple, fast, linear boundary, lower variance
  cons: pron to underfitting, lower accuracy

- Random Forest

  pro: high capacity, high accuracy, robust to overfitting, model explainability, no need for feature scaling, low bias and low variance
  cons: training slow, not as good on unbalanced data as XGBoost.

- XGBoost

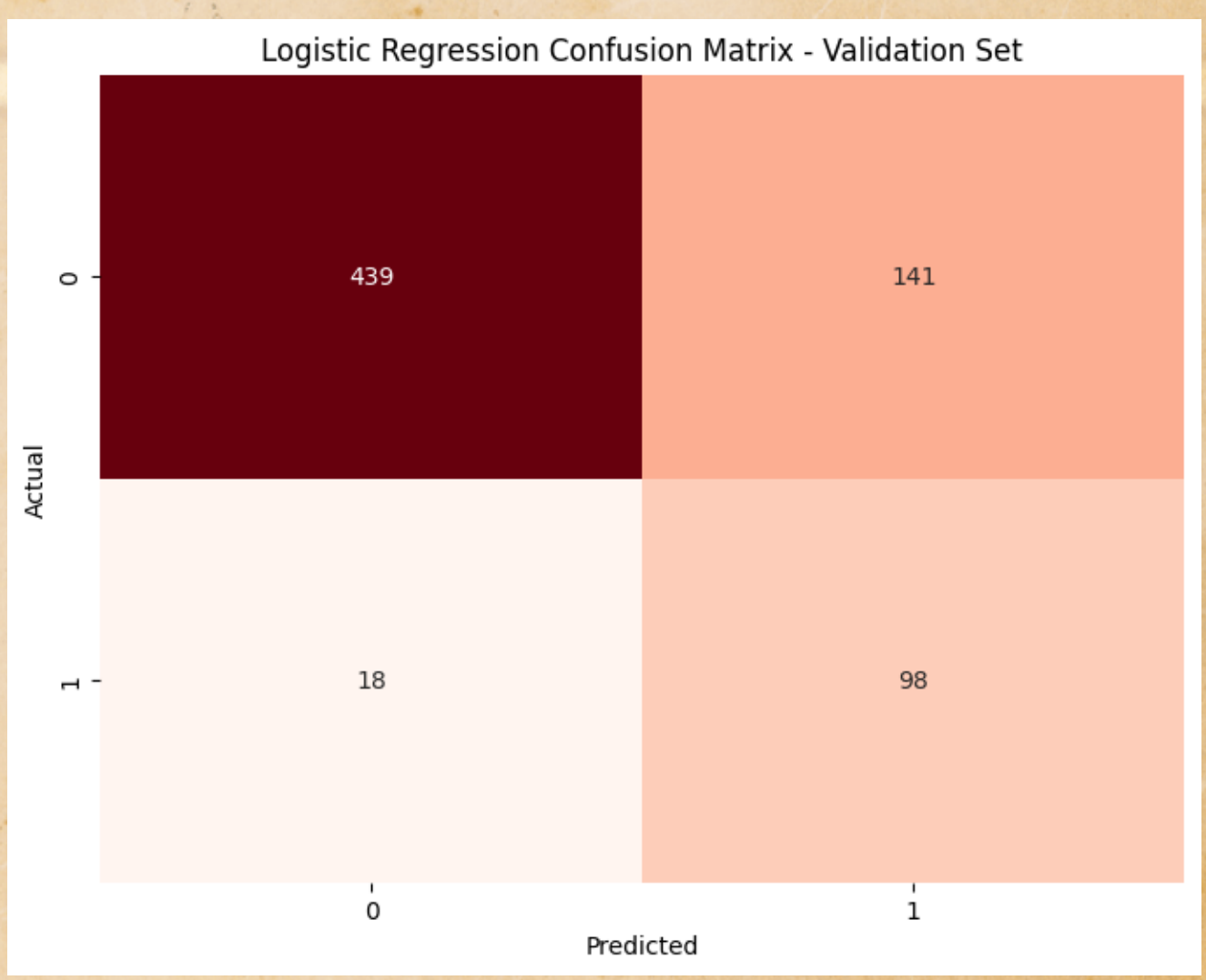
  Pros: usually with highest accuracy of the 3, very good at unbalanced data, distributed computation, low bias and low variance

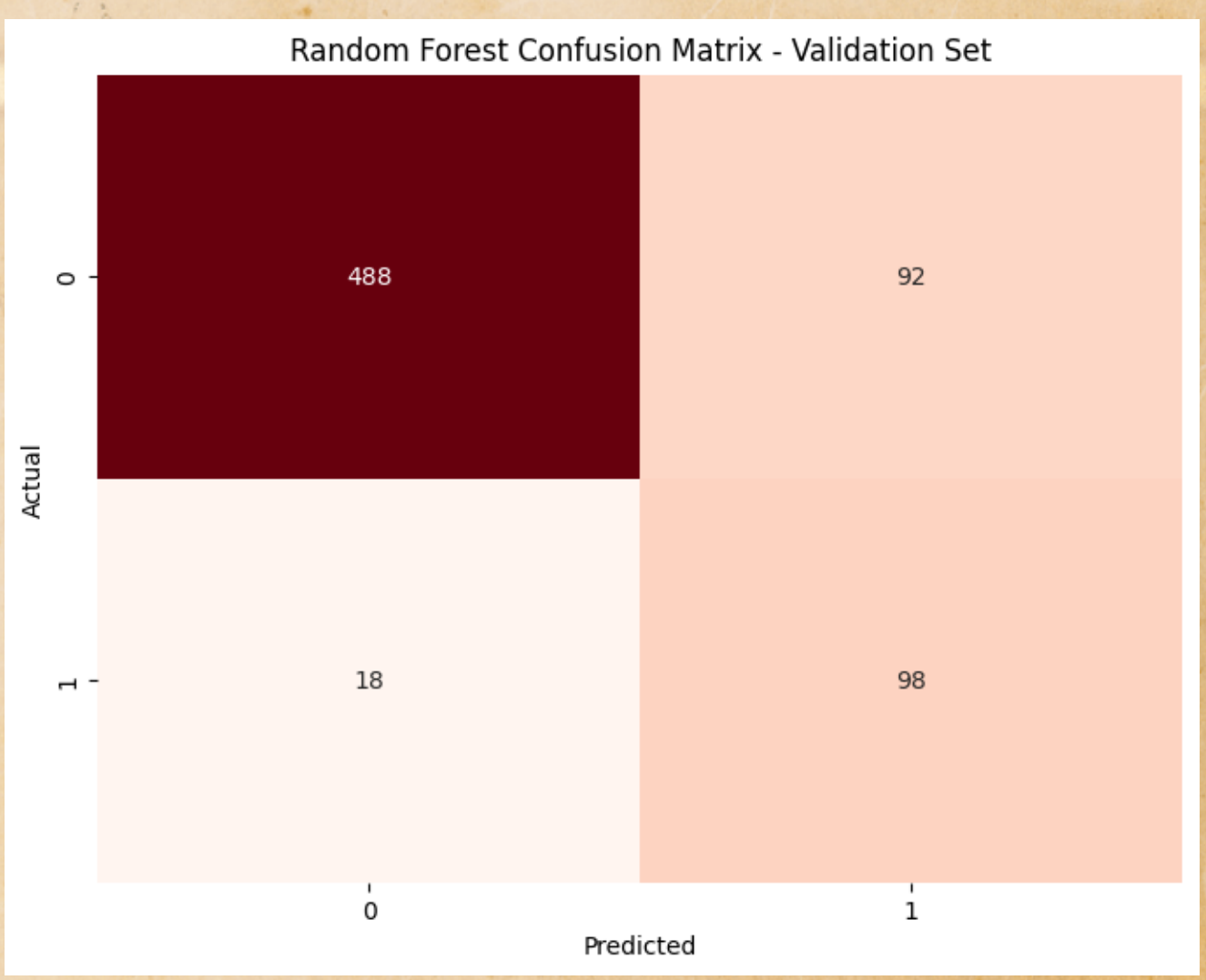
Evaluation metrics

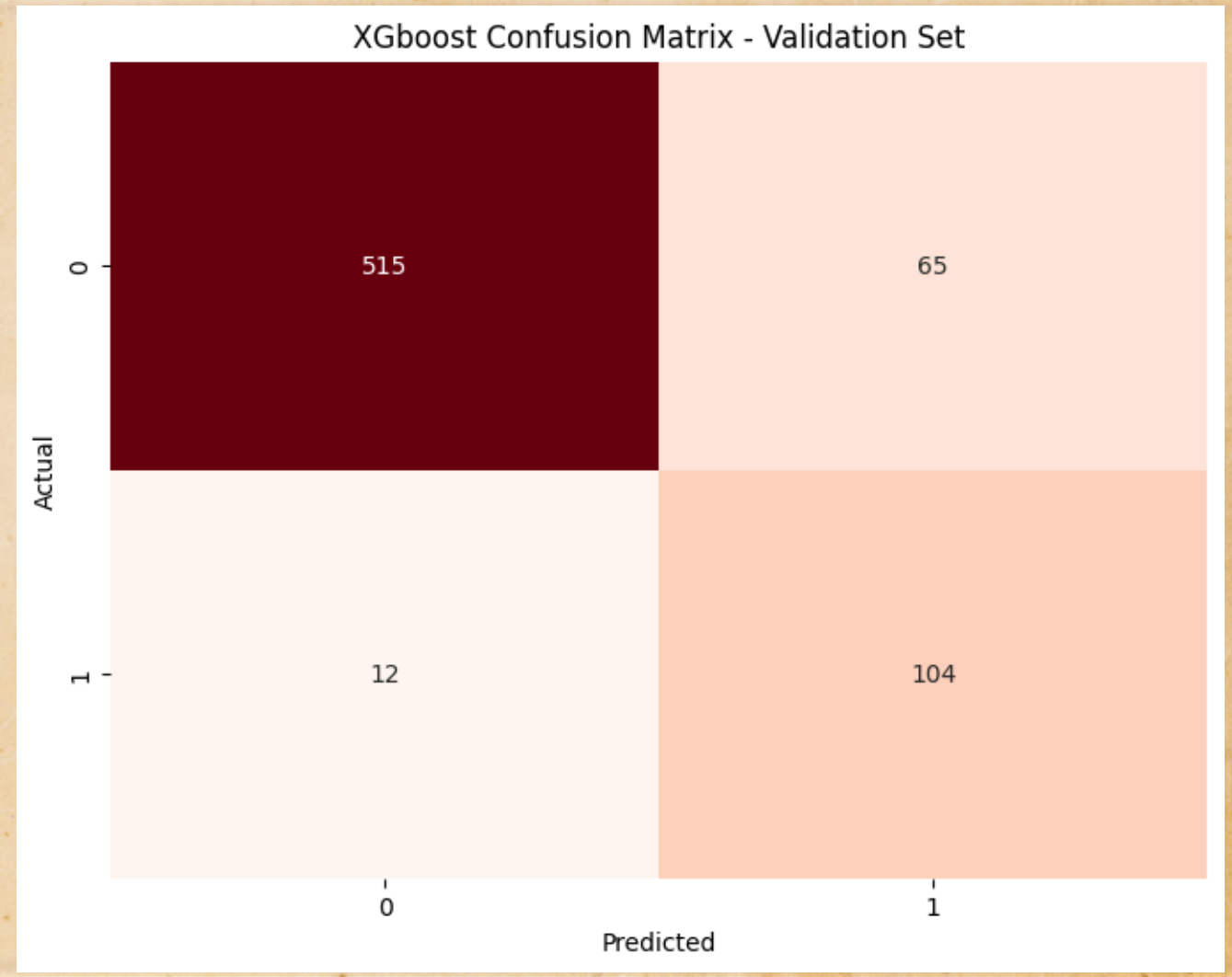
- Accuracy
- Precision
- Recall
- F1
- ROC AUC Score
- Confusion Matrix
-

Evaluation Matrics

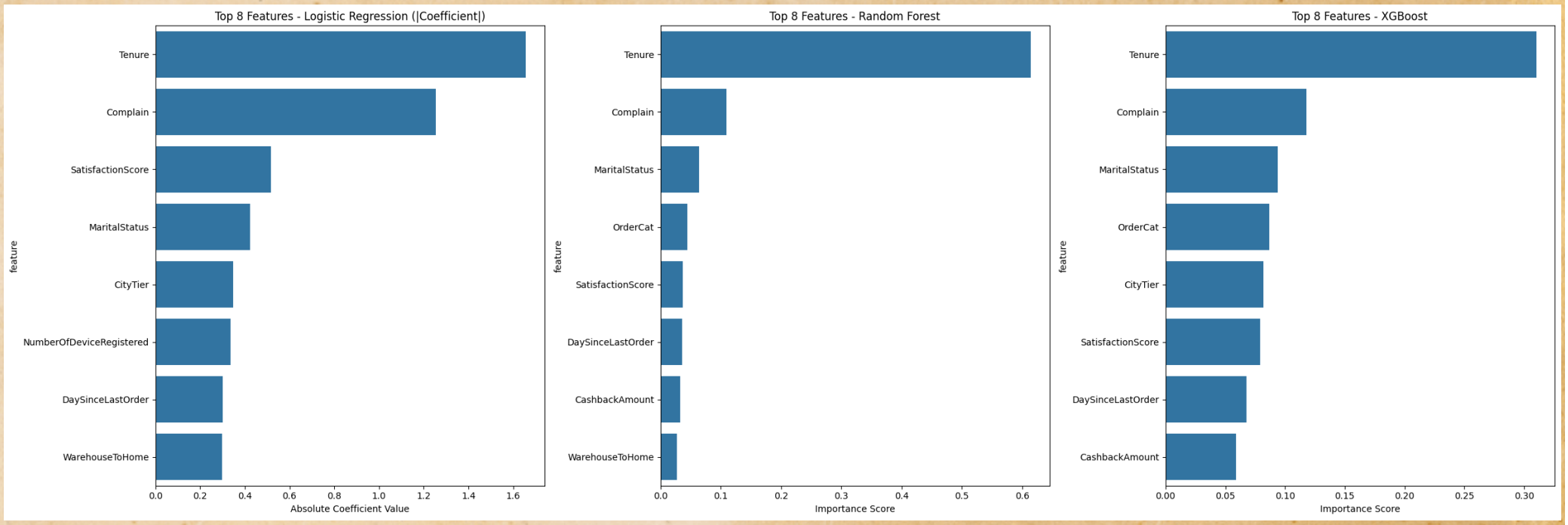
Evaluation set metrics	Logistic Regression	Random Forest	XGBoost
Accuracy	0.7716	0.8420	0.8894
ROC AUC Score	0.8735	0.9146	0.9536
Precision	0.41	0.52	0.62
Recall	0.84	0.84	0.90
F1-score	0.55	0.64	0.73


Logistic Regression Confusion Matrix - Validation Set
439
141
18
98
0
1
Actual
0
1
Predicted


Random Forest Confusion Matrix - Validation Set
488
92
18
98
Actual
0
1
0
1
Predicted


XGboost Confusion Matrix - Validation Set
515
65
12
104
0
1
Actual
0
1
Predicted

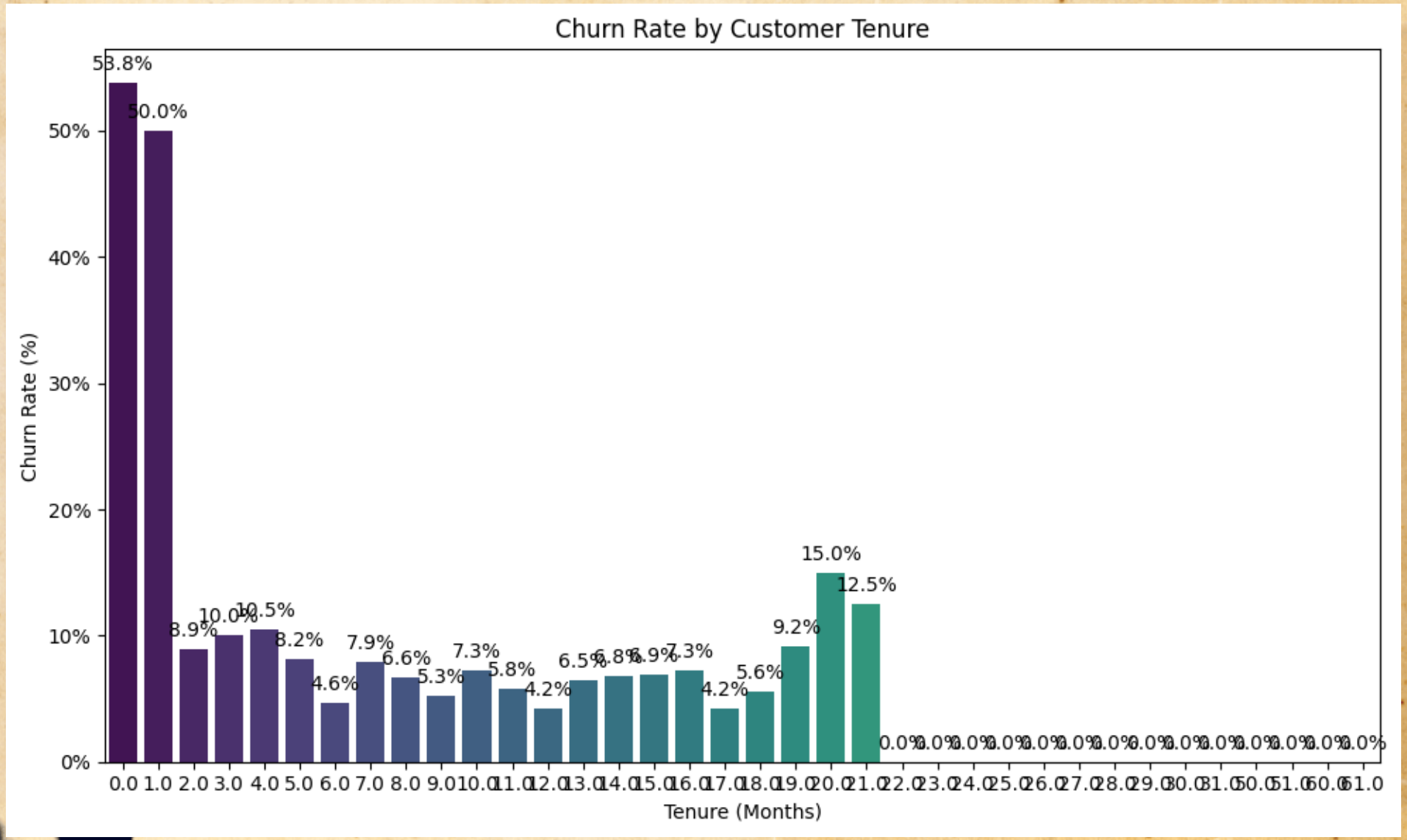
Feature Importance Analysis

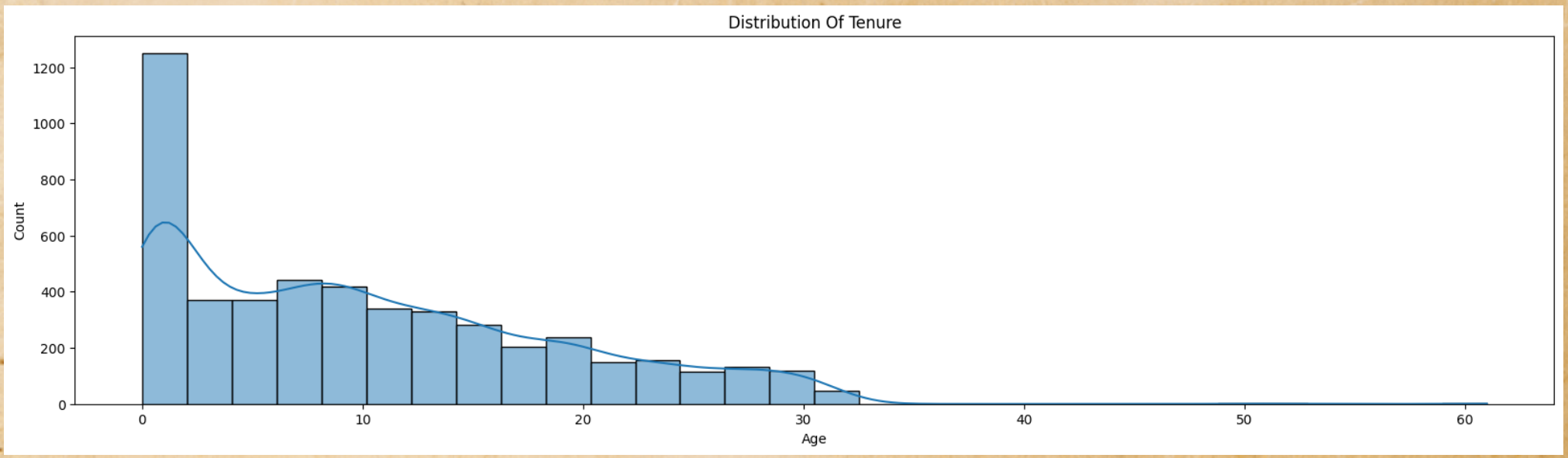


Evaluation – 5-Fold Cross-Validation

- Logistic Regression
  - Accuracy: 0.7608 (+/- 0.0119)
  - ROC AUC: 0.8491 (+/- 0.0253)
  - Average Precision: 0.6492 (+/- 0.0452)
- Random Forest
  - Accuracy: 0.8344 (+/- 0.0142)
  - ROC AUC: 0.8969 (+/- 0.0255)
  - Average Precision: 0.7028 (+/- 0.0425)
- XGBoost
  - Accuracy: 0.9080 (+/- 0.0097)
  - ROC AUC: 0.9498 (+/- 0.0173)
  - Average Precision: 0.8137 (+/- 0.0383)

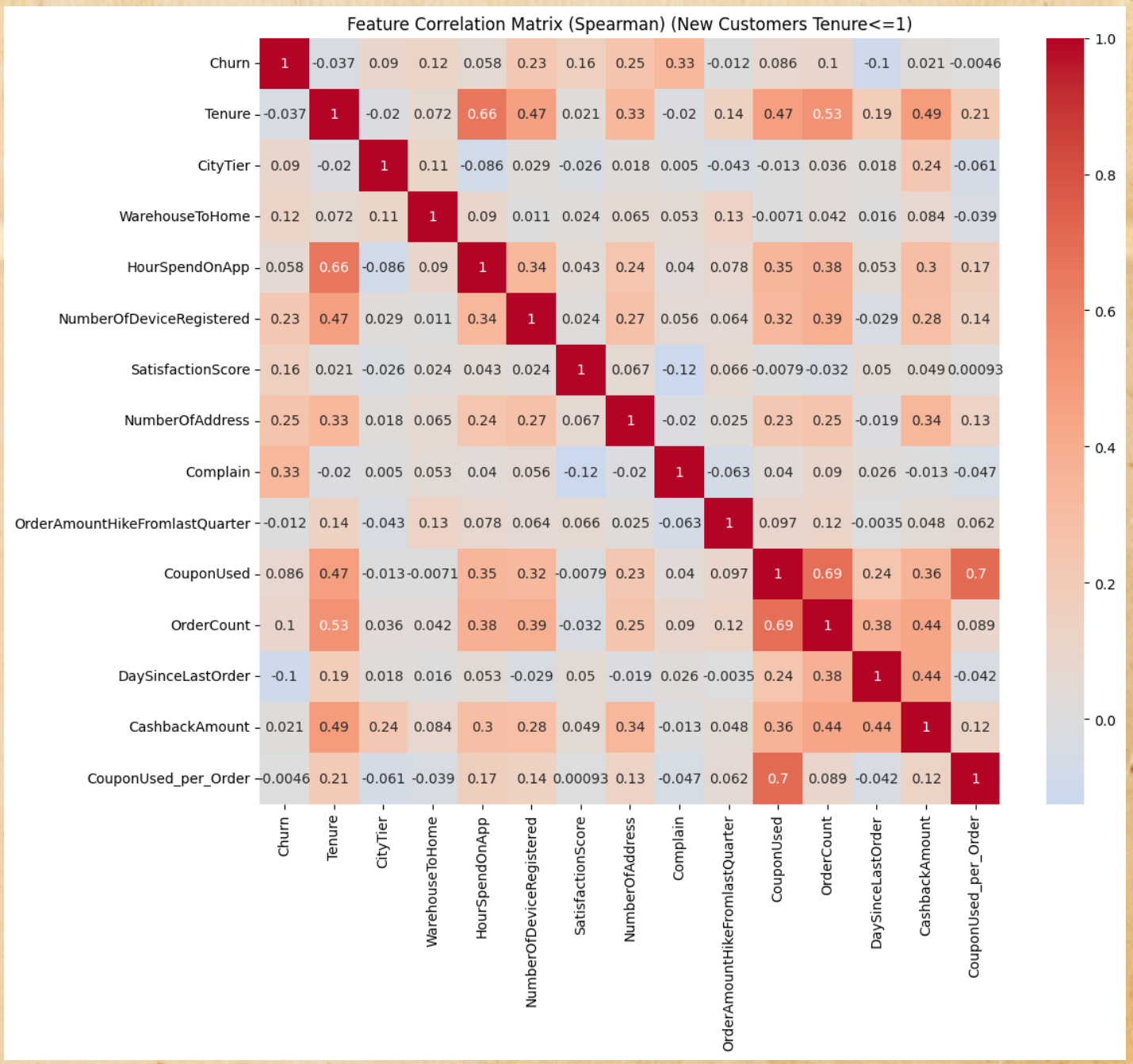
1. More Tenure Analysis & Related Predict Analysis

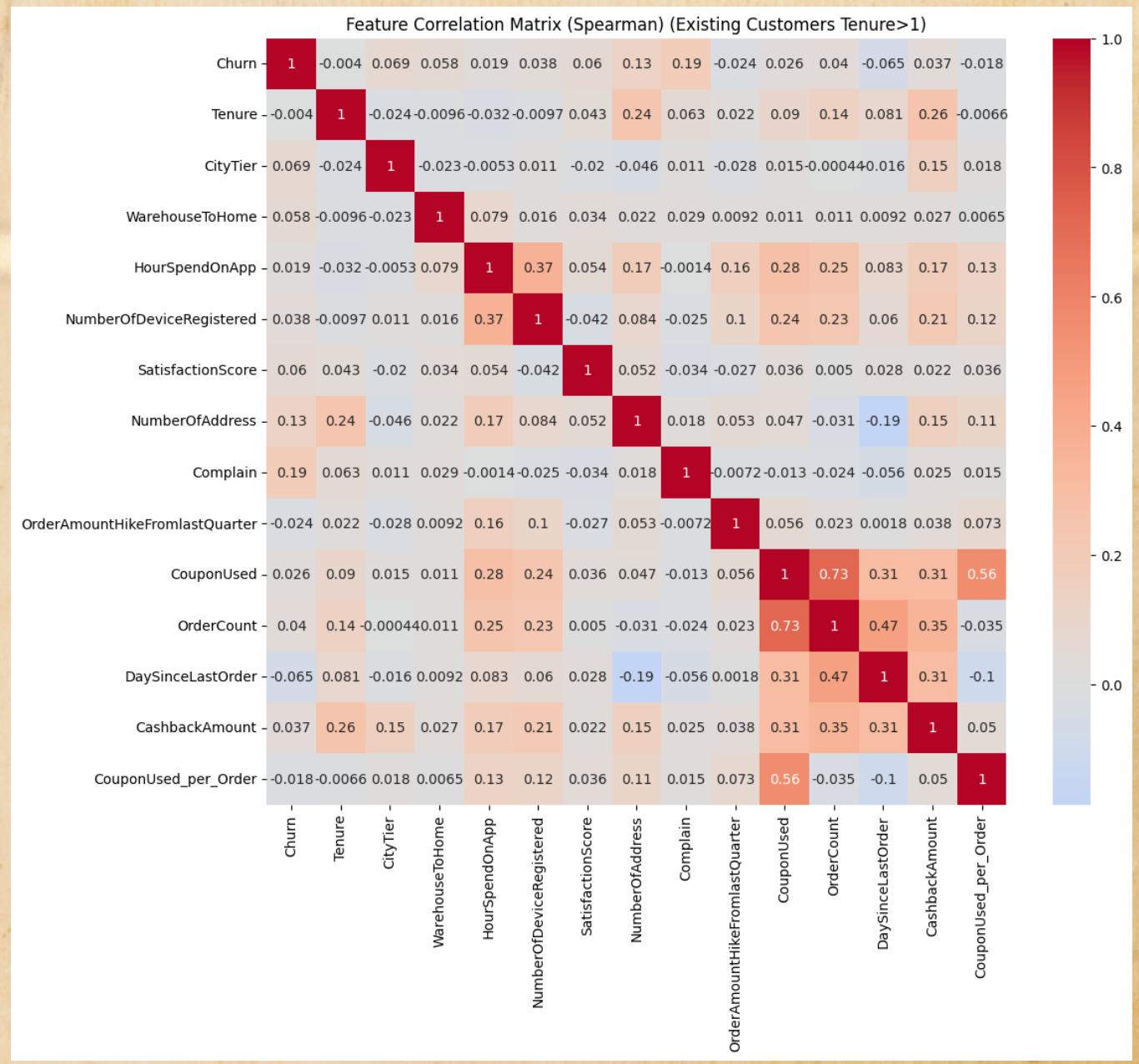


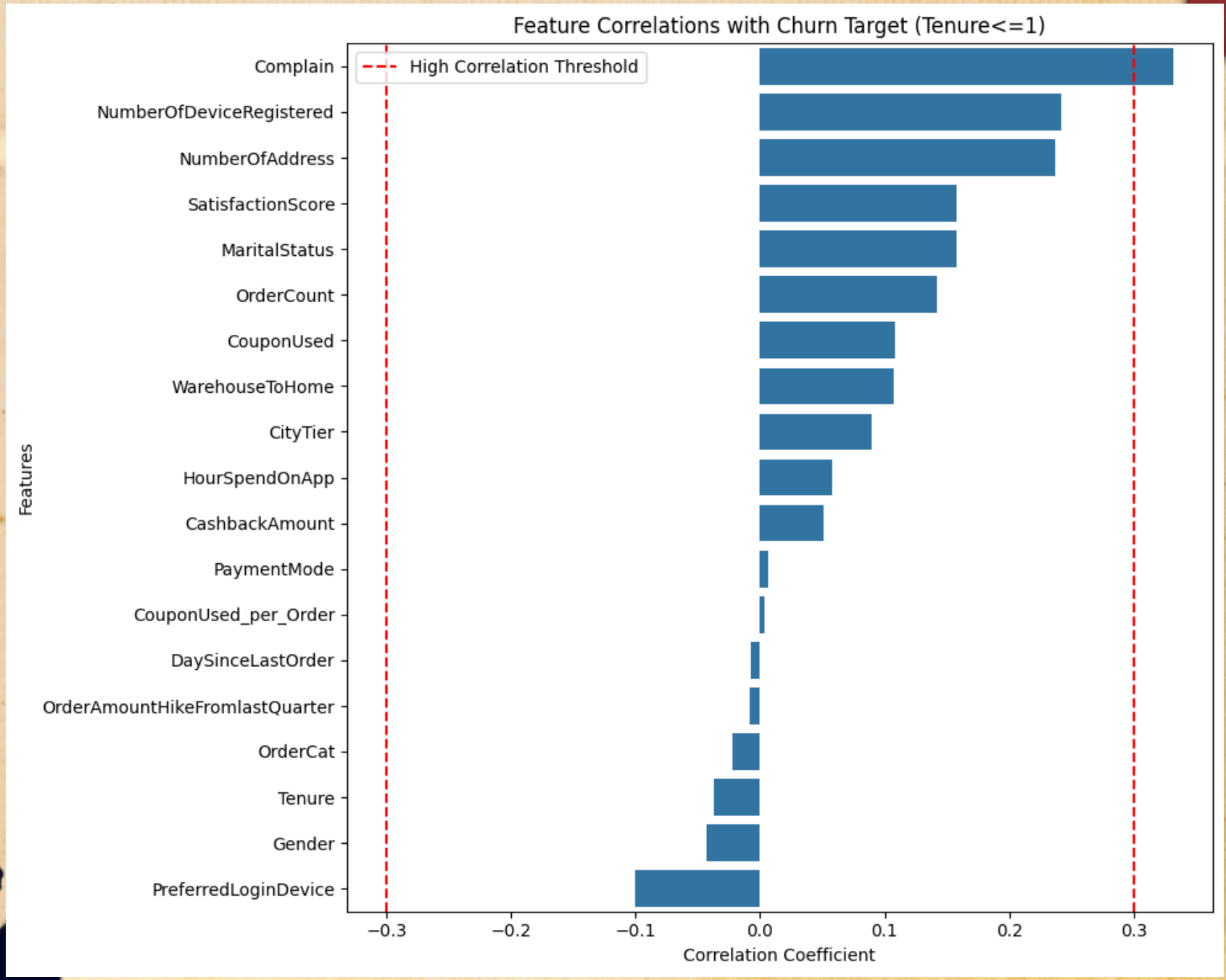

Distribution Of Tenure
Count
Age
0
200
400
600
800
1000
1200
0
10
20
30
40
50
60

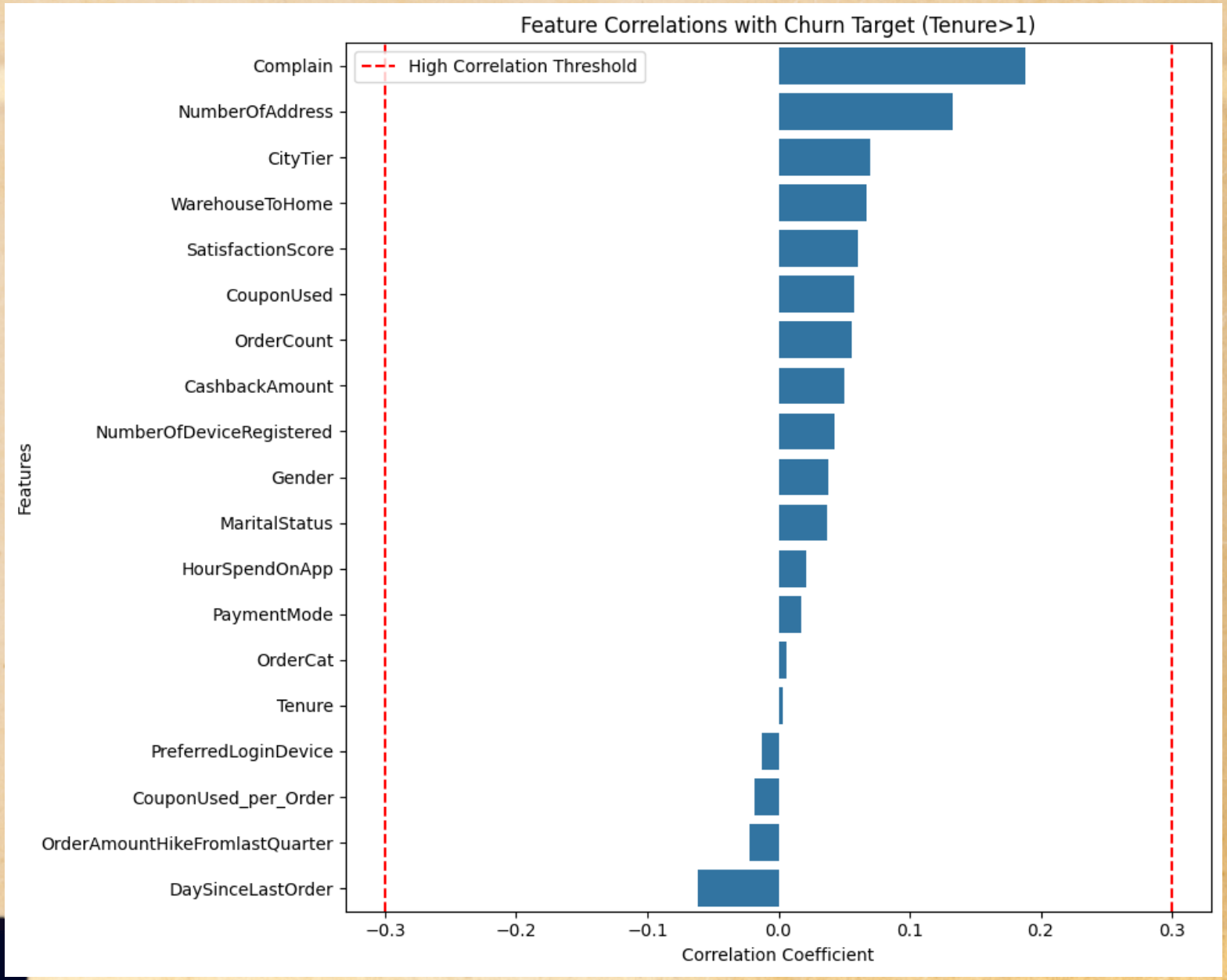
2 models

- 2 dataset
- Tenure<=1
- Tenure>1


Feature Correlation Matrix (Spearman) (New Customers Tenure<=1)
Churn 1 -0.037 0.09 0.12 0.058 0.23 0.16 0.25 0.33 -0.012 0.086 0.1 -0.1 0.021 -0.0046
Tenure -0.037 1 -0.02 0.072 0.66 0.47 0.021 0.33 -0.02 0.14 0.47 0.53 0.19 0.49 0.21
CityTier 0.09 -0.02 1 0.11 -0.086 0.029 -0.026 0.018 0.005 -0.043 -0.013 0.036 0.018 0.24 -0.061
WarehouseToHome 0.12 0.072 0.11 1 0.09 0.011 0.024 0.065 0.053 0.13 -0.0071 0.042 0.016 0.084 -0.039
HourSpendOnApp 0.058 0.66 -0.086 0.09 1 0.34 0.043 0.24 0.04 0.078 0.35 0.38 0.053 0.3 0.17
NumberOfDeviceRegistered 0.23 0.47 0.029 0.011 0.34 1 0.024 0.27 0.056 0.064 0.32 0.39 -0.029 0.28 0.14
SatisfactionScore 0.16 0.021 -0.026 0.024 0.043 0.024 1 0.067 -0.12 0.066 -0.0079 -0.032 0.05 0.049 0.00093
NumberOfAddress 0.25 0.33 0.018 0.065 0.24 0.27 0.067 1 -0.02 0.025 0.23 0.25 -0.019 0.34 0.13
Complain 0.33 -0.02 0.005 0.053 0.04 0.056 -0.12 -0.02 1 -0.063 0.04 0.09 0.026 -0.013 -0.047
OrderAmountHikeFromlastQuarter -0.012 0.14 -0.043 0.13 0.078 0.064 0.066 0.025 -0.063 1 0.097 0.12 -0.0035 0.048 0.062
CouponUsed 0.086 0.47 -0.013 -0.0071 0.35 0.32 -0.0079 0.23 0.04 0.097 1 0.69 0.24 0.36 0.7
OrderCount 0.1 0.53 0.036 0.042 0.38 0.39 -0.032 0.25 0.09 0.12 0.69 1 0.38 0.44 0.089
DaySinceLastOrder -0.1 0.19 0.018 0.016 0.053 -0.029 0.05 -0.019 0.026 -0.0035 0.24 0.38 1 0.44 -0.042
CashbackAmount 0.021 0.49 0.24 0.084 0.3 0.28 0.049 0.34 -0.013 0.048 0.36 0.44 0.44 1 0.12
CouponUsed_per_Order -0.0046 0.21 -0.061 -0.039 0.17 0.14 0.00093 0.13 -0.047 0.062 0.7 0.089 -0.042 0.12 1
Churn
Tenure
CityTier
WarehouseToHome
HourSpendOnApp
NumberOfDeviceRegistered
SatisfactionScore
NumberOfAddress
Complain
OrderAmountHikeFromlastQuarter
CouponUsed
OrderCount
DaySinceLastOrder
CashbackAmount
CouponUsed_per_Order
1.0
0.8
0.6
0.4
0.2
0.0


Feature Correlation Matrix (Spearman) (Existing Customers Tenure>1)
Churn 1 -0.004 0.069 0.058 0.019 0.038 0.06 0.13 0.19 -0.024 0.026 0.04 -0.065 0.037 -0.018
Tenure -0.004 1 -0.024 -0.0096 -0.032 -0.0097 0.043 0.24 0.063 0.022 0.09 0.14 0.081 0.26 -0.0066
CityTier 0.069 -0.024 1 -0.023 -0.0053 0.011 -0.02 -0.046 0.011 -0.028 0.015 -0.00044 -0.016 0.15 0.018
WarehouseToHome 0.058 -0.0096 -0.023 1 0.079 0.016 0.034 0.022 0.029 0.0092 0.011 0.011 0.0092 0.027 0.0065
HourSpendOnApp 0.019 -0.032 -0.0053 0.079 1 0.37 0.054 0.17 -0.0014 0.16 0.28 0.25 0.083 0.17 0.13
NumberOfDeviceRegistered 0.038 -0.0097 0.011 0.016 0.37 1 -0.042 0.084 -0.025 0.1 0.24 0.23 0.06 0.21 0.12
SatisfactionScore 0.06 0.043 -0.02 0.034 0.054 -0.042 1 0.052 -0.034 -0.027 0.036 0.005 0.028 0.022 0.036
NumberOfAddress 0.13 0.24 -0.046 0.022 0.17 0.084 0.052 1 0.018 0.053 0.047 -0.031 -0.19 0.15 0.11
Complain 0.19 0.063 0.011 0.029 -0.0014 -0.025 -0.034 0.018 1 -0.0072 -0.013 -0.024 -0.056 0.025 0.015
OrderAmountHikeFromlastQuarter -0.024 0.022 -0.028 0.0092 0.16 0.1 -0.027 0.053 -0.0072 1 0.056 0.023 0.0018 0.038 0.073
CouponUsed 0.026 0.09 0.015 0.011 0.28 0.24 0.036 0.047 -0.013 0.056 1 0.73 0.31 0.31 0.56
OrderCount 0.04 0.14 -0.00044 0.011 0.25 0.23 0.005 -0.031 -0.024 0.023 0.73 1 0.47 0.35 -0.035
DaySinceLastOrder -0.065 0.081 -0.016 0.0092 0.083 0.06 0.028 -0.19 -0.056 0.0018 0.31 0.47 1 0.31 -0.1
CashbackAmount 0.037 0.26 0.15 0.027 0.17 0.21 0.022 0.15 0.025 0.038 0.31 0.35 0.31 1 0.05
CouponUsed_per_Order -0.018 -0.0066 0.018 0.0065 0.13 0.12 0.036 0.11 0.015 0.073 0.56 -0.035 -0.1 0.05 1
Churn
Tenure
CityTier
WarehouseToHome
HourSpendOnApp
NumberOfDeviceRegistered
SatisfactionScore
NumberOfAddress
Complain
OrderAmountHikeFromlastQuarter
CouponUsed
OrderCount
DaySinceLastOrder
CashbackAmount
CouponUsed_per_Order
1.0
0.8
0.6
0.4
0.2
0.0


Feature Correlations with Churn Target (Tenure<=1)
High Correlation Threshold
Complain
NumberOfDeviceRegistered
NumberOfAddress
SatisfactionScore
MaritalStatus
OrderCount
CouponUsed
WarehouseToHome
CityTier
HourSpendOnApp
CashbackAmount
PaymentMode
CouponUsed_per_Order
DaySinceLastOrder
OrderAmountHikeFromlastQuarter
OrderCat
Tenure
Gender
PreferredLoginDevice
Features
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Correlation Coefficient


Feature Correlations with Churn Target (Tenure>1)
High Correlation Threshold
Complain
NumberOfAddress
CityTier
WarehouseToHome
SatisfactionScore
CouponUsed
OrderCount
CashbackAmount
NumberOfDeviceRegistered
Gender
MaritalStatus
HourSpendOnApp
PaymentMode
OrderCat
Tenure
PreferredLoginDevice
CouponUsed_per_Order
OrderAmountHikeFromlastQuarter
DaySinceLastOrder
Features
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Correlation Coefficient

2. Actionable Inside and recommendations

- Churn prediction with probability for customer retention
- Different customer retention strategy for new customers(tenure<=1) and existing ones(tenure>1)
- Consider remove sensitive features like Sex and MaritalStatus for AI data regulation related DEI
- Collect data with actionable features like sales/marketing related ones, promotion by email, samples sent, visual/phone communication logs… which can be experimented to lower the churn

Questions?

Thanks