Predicting E - Commerce Churn

Isaac Obodo
Tapleen Lamba
Ashan Dias

Churn?!

The rate at which customers leave your company

Importance?!

Lost customers, lost revenue!!!

It’s a lot more expensive to attract new customers rather than retaining the existing ones

Introduction and Problem Statement

Brief Overview of the Project & Customer Churn

Project Overview:

- Objective: Implement machine learning strategies to predict customer churn in e-commerce.
- Methodology: Employ logistic regression and neural network models for analysis.
- Data Utilized: Analysis leverages an e-commerce dataset, including customer preferences, satisfaction, and transactional history.

DATA WE HAVE

DATA DICTIONARY

Data	Variable	Discerption
E Comm	CustomerID	Unique customer ID
E Comm	Churn	Churn Flag
E Comm	Tenure	Tenure of customer in organization
E Comm	PreferredLoginDevice	Preferred login device of customer
E Comm	CityTier	City tier
E Comm	WarehouseToHome	Distance in between warehouse to home of customer
E Comm	PreferredPaymentMode	Preferred payment method of customer
E Comm	Gender	Gender of customer
E Comm	HourSpendOnApp	Number of hours spend on mobile application or website
E Comm	NumberOfDeviceRegistered	Total number of deceives is registered on particular customer
E Comm	PreferedOrderCat	Preferred order category of customer in last month
E Comm	SatisfactionScore	Satisfactory score of customer on service
E Comm	MaritalStatus	Marital status of customer
E Comm	NumberOfAddress	Total number of added added on particular customer
E Comm	Complain	Any complaint has been raised in last month
E Comm	OrderAmountHikeFromlastYear	Percentage increases in order from last year
E Comm	CouponUsed	Total number of coupon has been used in last month
E Comm	OrderCount	Total number of orders has been places in last month
E Comm	DaySinceLastOrder	Day Since last order by customer
E Comm	CashbackAmount	Average cashback in last month

Dataset and Preprocessing

Dataset

- Content: Contains detailed customer information, including preferences, satisfaction scores, and transaction data.
- Features: Includes variables like CustomerID, Churn, Tenure, PreferredLoginDevice, CityTier, and more, totaling over 15 distinct attributes.
- Purpose: Designed to offer insights into customer behavior and factors influencing churn.

One-Hot Encoding

Converted categorical variables (e.g., PreferredLoginDevice, Gender, MaritalStatus) into a numerical format suitable for machine learning models.

Class Imbalance Addressing and Scaling

Implemented oversampling on the training data to counteract class imbalance in the target variable (Churn)

Used the MinMax Scaler to scale values.

Handling Missing Values

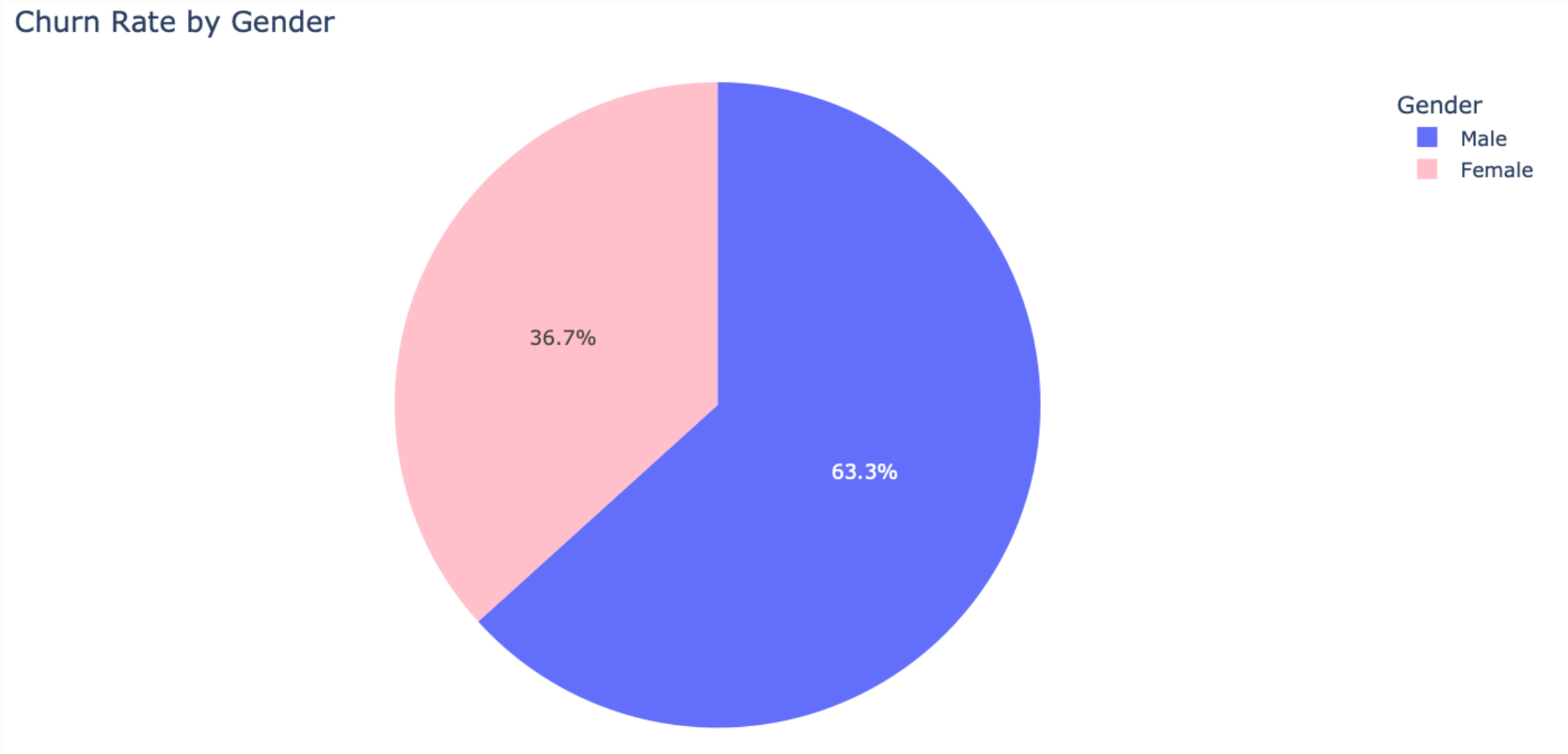
Applied iterative imputation techniques to fill in missing data, ensuring dataset integrity.

Data Splitting

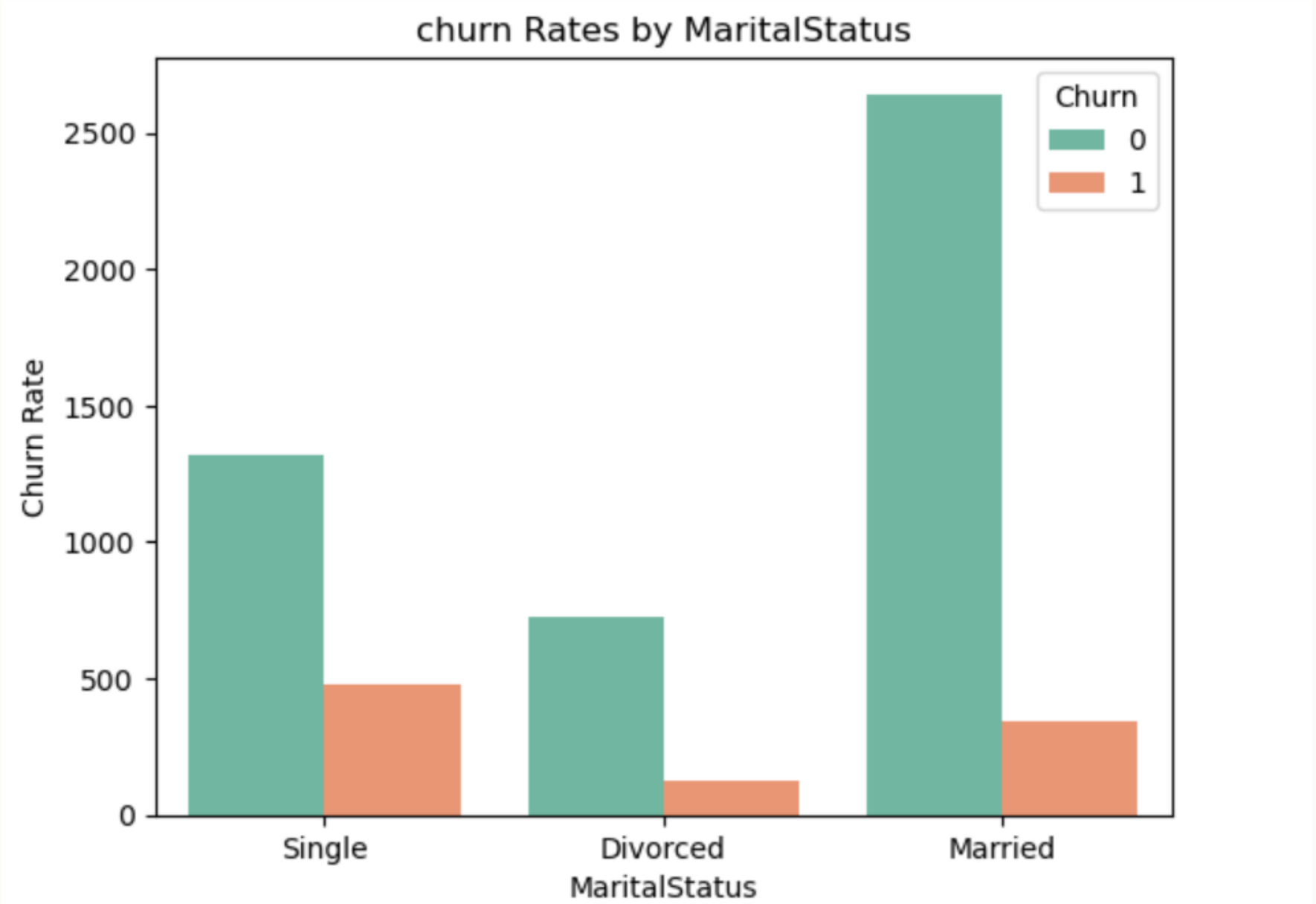
Segregated the dataset into training (75%) and testing (25%) sets to validate model performance.


Navigating the Data
Galaxy with Matplotlib
and Pandas!

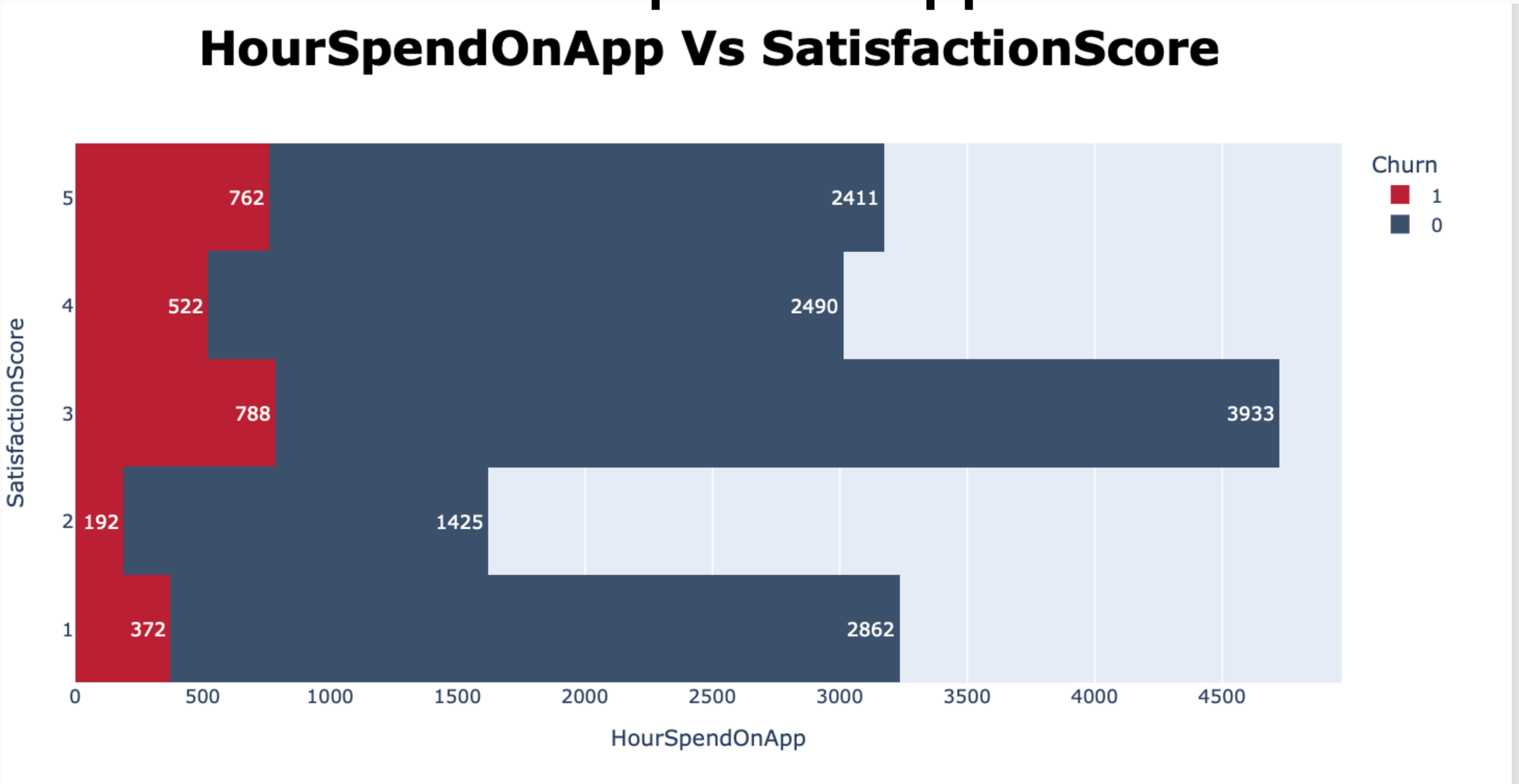
RELATIONSHIP BETWEEN GENDER AND CHURN



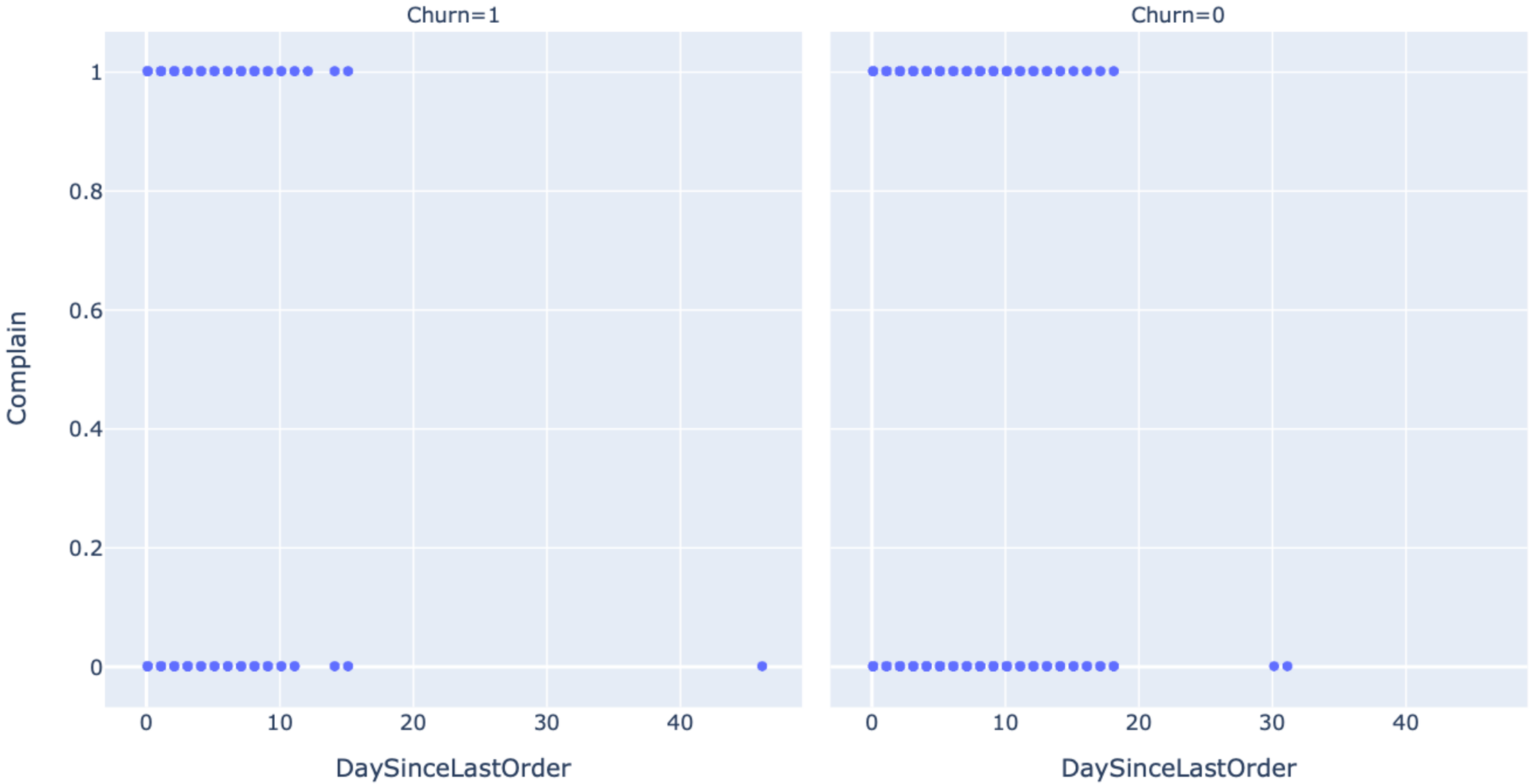
MARITAL STATUS AND THE CHURN RATE



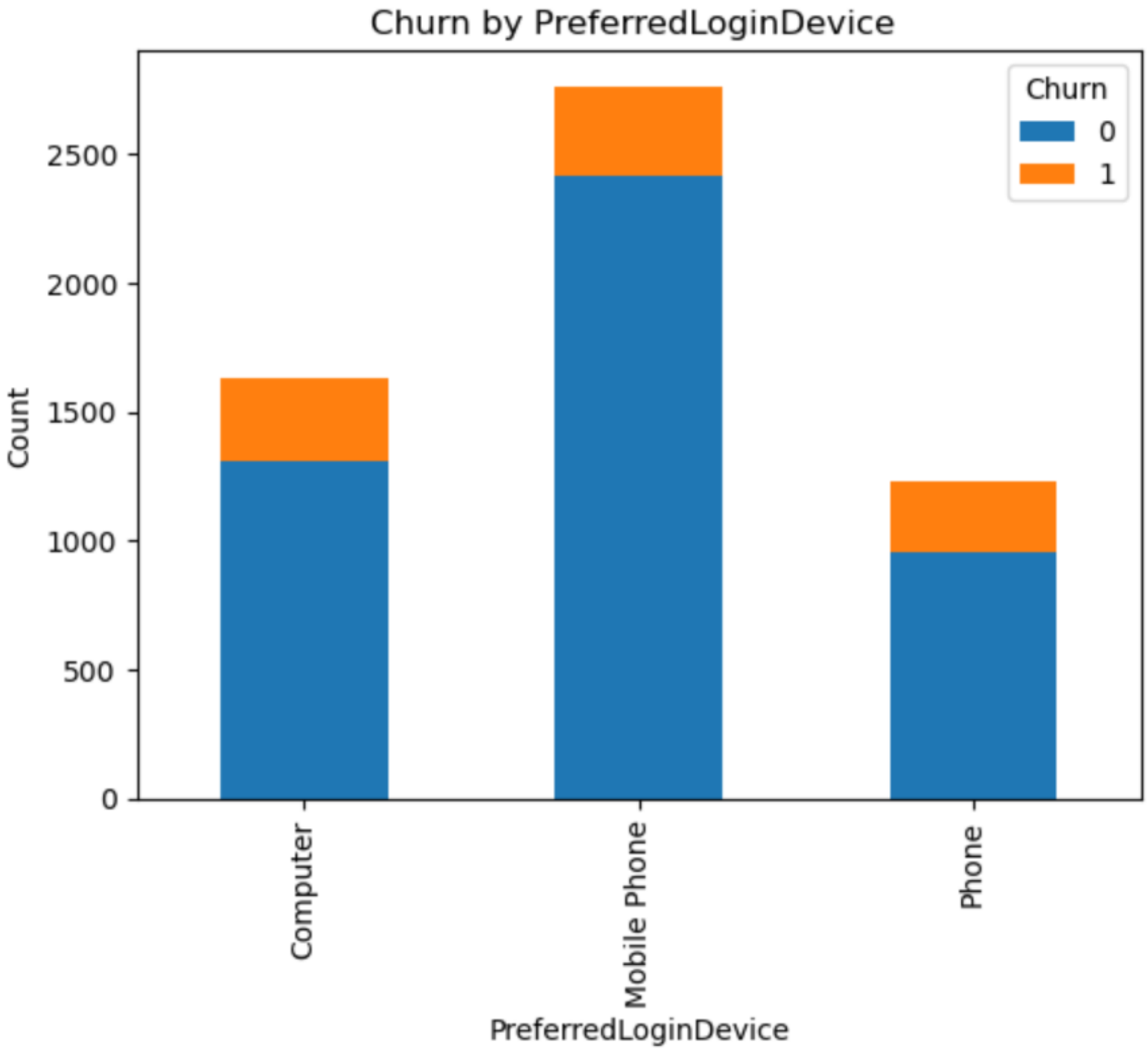
Is there a correlation between SatisfactionScore and HourSpendOnApp?



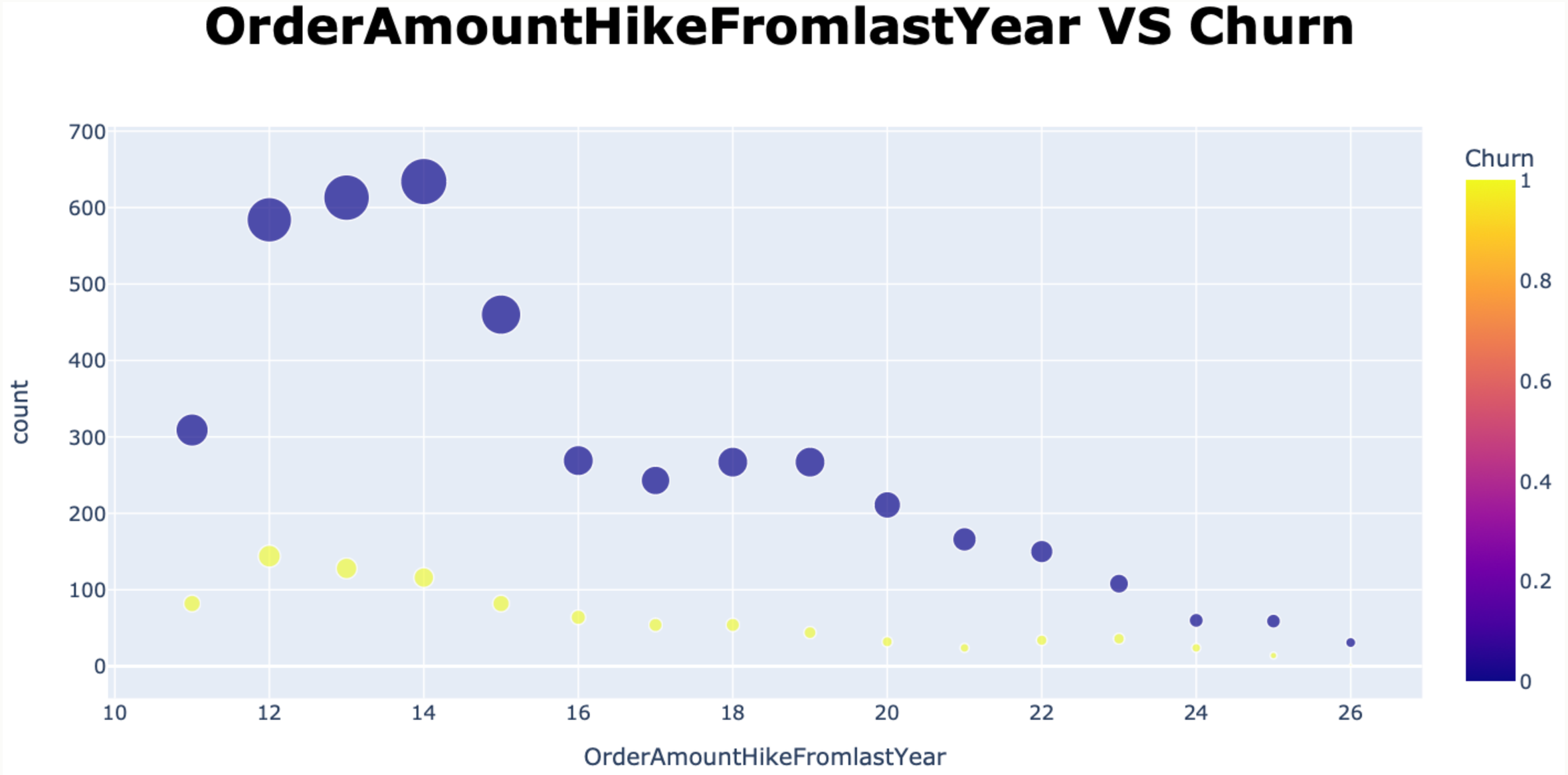
COMPLAIN AND DAY SINCE LAST ORDERED



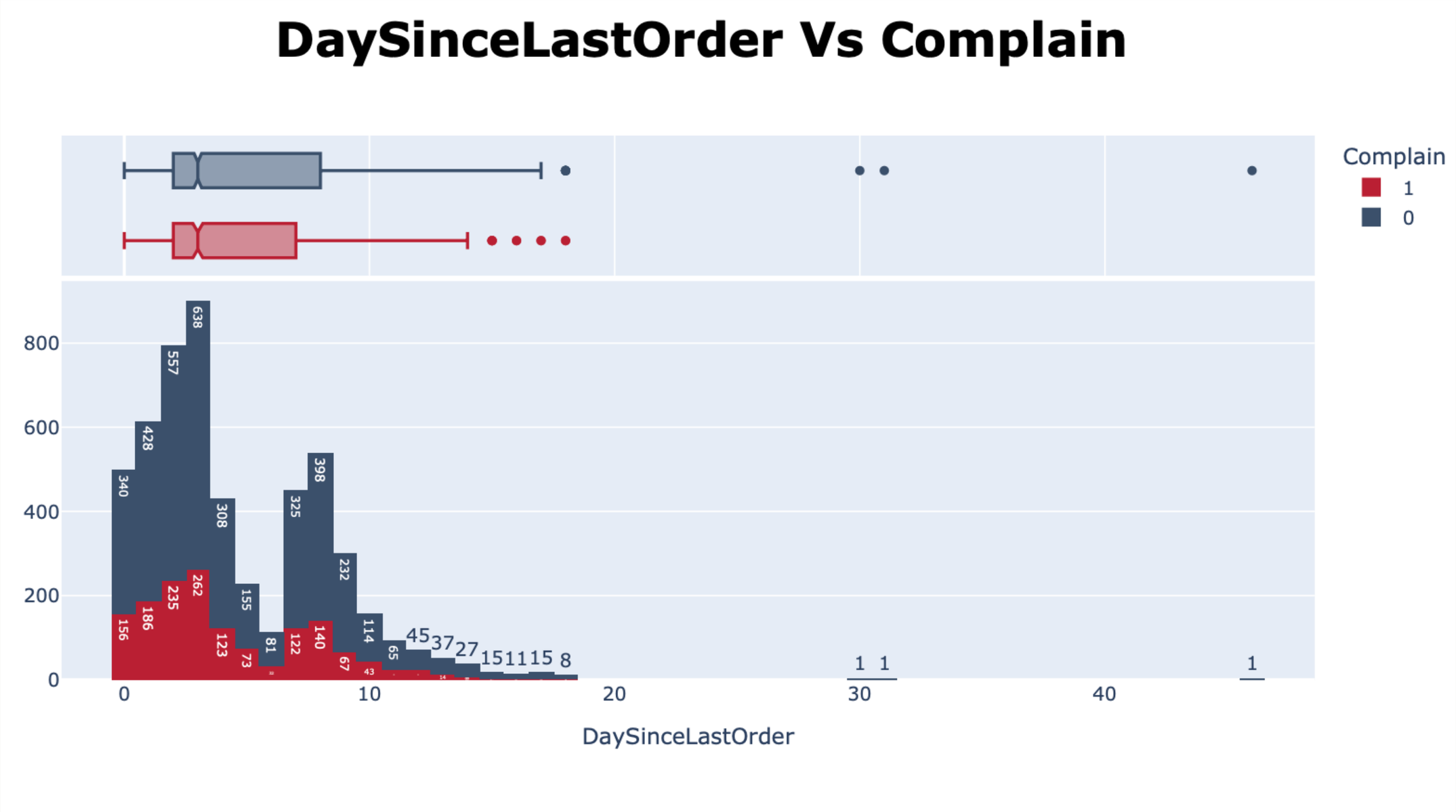
PREFERRED LOGIN AND CHURN



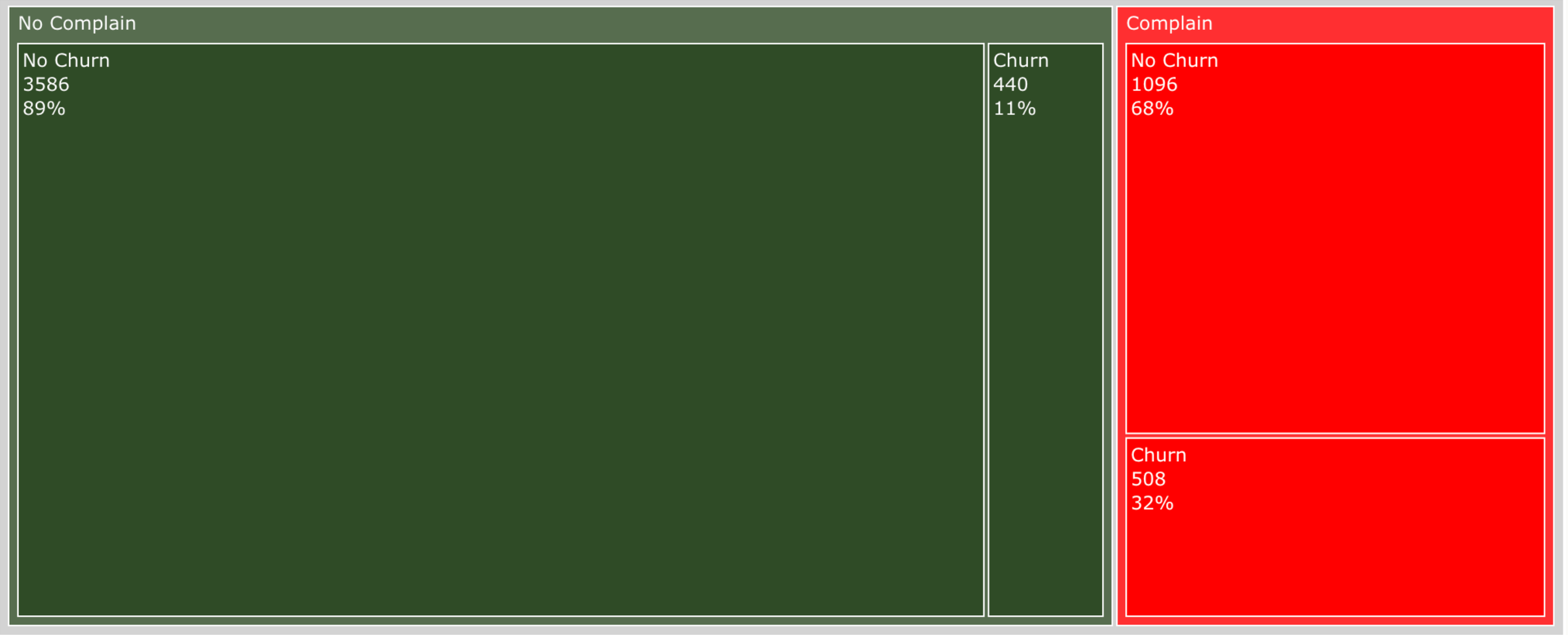
Does the percentage increase in order amount from last year affect churn rate?



What is the relation between Complain and DaySinceLastOrder for churned customers?



Are customers who complained more likely to churn?



Model Development

Logistic Regression

A statistical model used for binary classification tasks. It predicts the probability of an event (churn) by fitting data to a logistic curve

Logistic regression is used to estimate the likelihood of a customer churning based on various predictors like tenure, satisfaction scores, etc.

Simple, interpretable, and efficient for linear relationships.

Logistic Regression Model Results

Key Metrics:

Accuracy: 88.78%

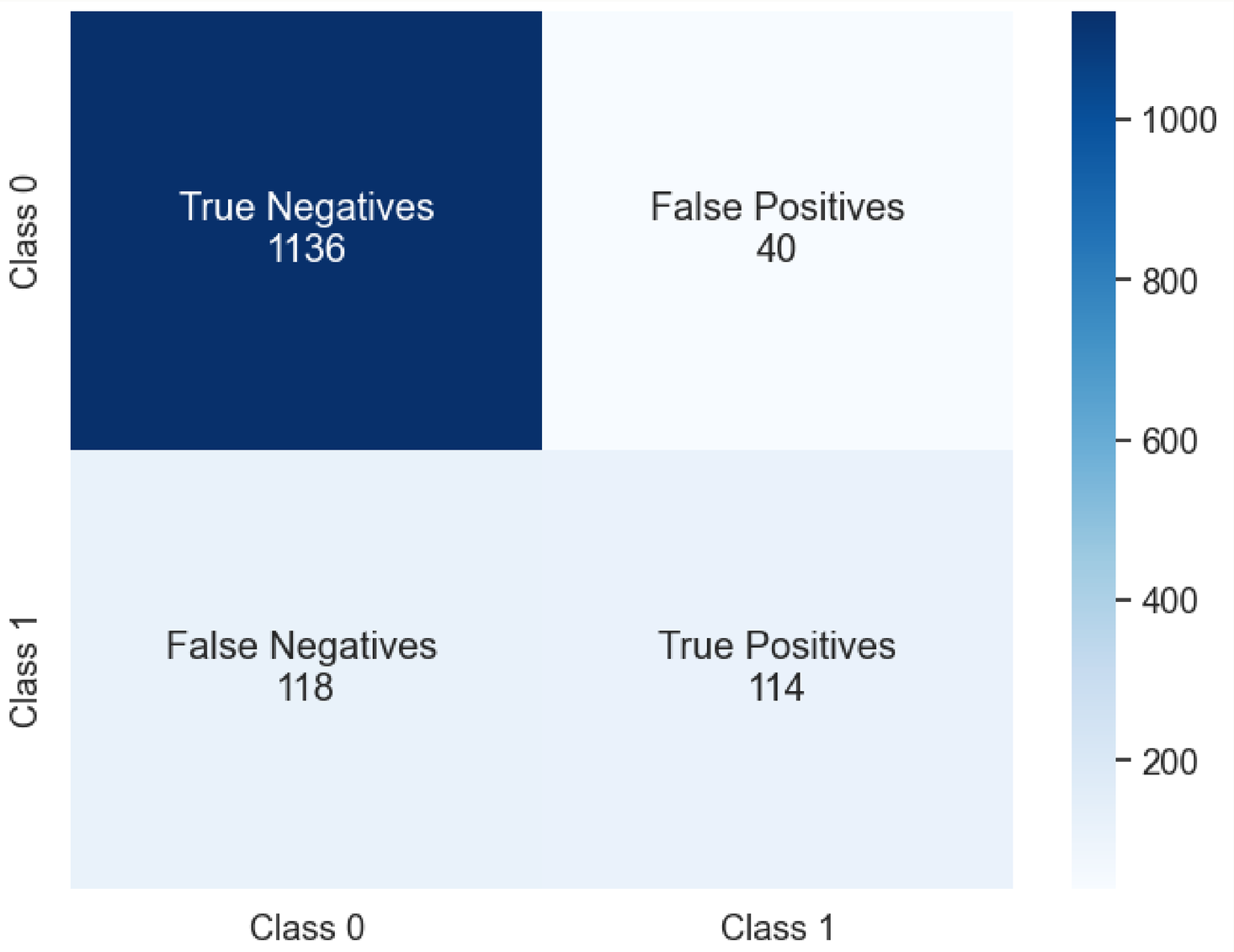
```
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.97      0.93      1176
           1       0.74      0.49      0.59       232

    accuracy                           0.89      1408
   macro avg       0.82      0.73      0.76      1408
weighted avg       0.88      0.89      0.88      1408

Confusion Matrix:
 [[1136   40]
 [ 118  114]]
```

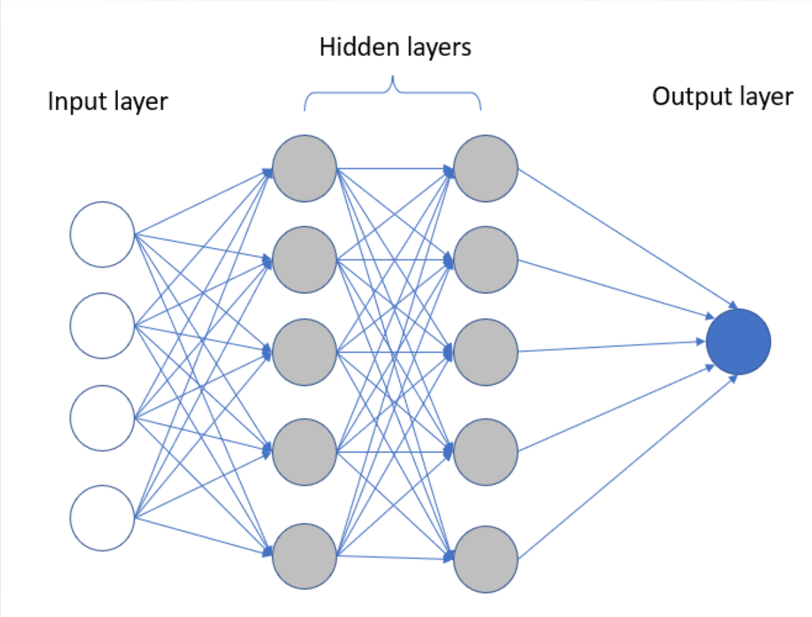
The confusion matrix further breaks down the model's performance



Model Development

Deep Neural Network

A more complex model that mimics the human brain's structure and function, using layers of nodes or 'neurons' to learn from data.



Implemented to capture non-linear relationships and interactions between variables in the dataset.

High accuracy, flexibility, and capability to model complex patterns

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 80)                2800

 dense_1 (Dense)             (None, 30)                2430

 dense_2 (Dense)             (None, 1)                 31

=================================================================
Total params: 5261 (20.55 KB)
Trainable params: 5261 (20.55 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Initial Deep Learning Model

Model Parameters

- Target variable: 'Churn'
- Features: All columns except 'Churn'

- 2 hidden Layers
  - Neurons in hidden layer 1 = 80
  - Neurons in hidden layer 2 = 30

- 1 output layer

- Activation Function: 'relu'

- Model Type: Sequential
  - Data flows sequentially from input layer to output layer

- Epochs = 50

Model was trained on training data and tested against testing data

Results:

Accuracy = 95.03%

Loss = 17.35%

```
Best val_accuracy So Far: 0.9829545617103577
Total elapsed time: 00h 15m 24s
```

```
44/44 - 0s - loss: 0.0861 - accuracy: 0.9830 - 173ms/epoch - 4ms/step
Loss: 0.08609770238399506, Accuracy: 0.9829545617103577
```

```
{'activation': 'tanh',
 'first_units': 81,
 'num_layers': 5,
 'units_0': 96,
 'units_1': 21,
 'units_2': 16,
 'units_3': 31,
 'units_4': 26,
 'units_5': 1,
 'tuner/epochs': 50,
 'tuner/initial_epoch': 0,
 'tuner/bracket': 0,
 'tuner/round': 0}
```

Optimized Deep Learning Model

Auto-Optimization Process

• Utilize the Keras-Tuner to get the best hyperparameters for the model by testing different combinations of parameters.

• Hyperparameter options set as variables for testing:
- Activation Function: ‘relu’, ‘tanh’, ‘sigmoid’
- Neurons in each hidden layer: min = 1, max = 100
- Number of hidden layers: 1 to 6

• Greater computing power can allow for even more fine-tuning of the model

Model was trained on training data and tested against testing data

Results:

Accuracy = 98.30%

Loss = 8.61%

Model Selection Rationale

- Complementarity: The combination of a simple, interpretable model (Logistic Regression) and a complex, high-performing one (Neural Network) provides a comprehensive analysis.

- Performance & Complexity Balance: Logistic Regression offers a baseline model, while the Neural Network explores deeper patterns in the data.

- Scope of Analysis: The two models together allow for a robust understanding of both straightforward and complex factors influencing customer churn.

Conclusions

The machine learning models developed in this project provide valuable insights into customer churn prediction for the e-commerce business. The logistic regression model achieved an accuracy of 88.78%, while the neural network model with hyperparameter optimization achieved an accuracy of 98.30%. The high accuracy of the neural network model suggests that it is a promising approach for customer churn prediction.