Enhancing Customer Churn Prediction: The Power of Statistical Feature Engineering

University of Southampton

Zhe GUAN

zg2u24@soton.ac.uk

Introduction

✳ The churn rate is now regarded as equally significant as financial profitability in assessing growth, and businesses try various methods to minimise churn to maintain customer retention.

✳Various statistical methods play an important role in enhancing model accuracy by transforming raw data into meaningful features that better capture patterns and trends.

✳ By analysing 1,000 consumer data from the Lloyds Bank on the Forage platform, the impact of statistical feature engineering techniques is validated by comparing customer churn predictions.

Methods

✳**Four different customer datasets** are analysed, including customer demographics, interactions, transactions and login activities.

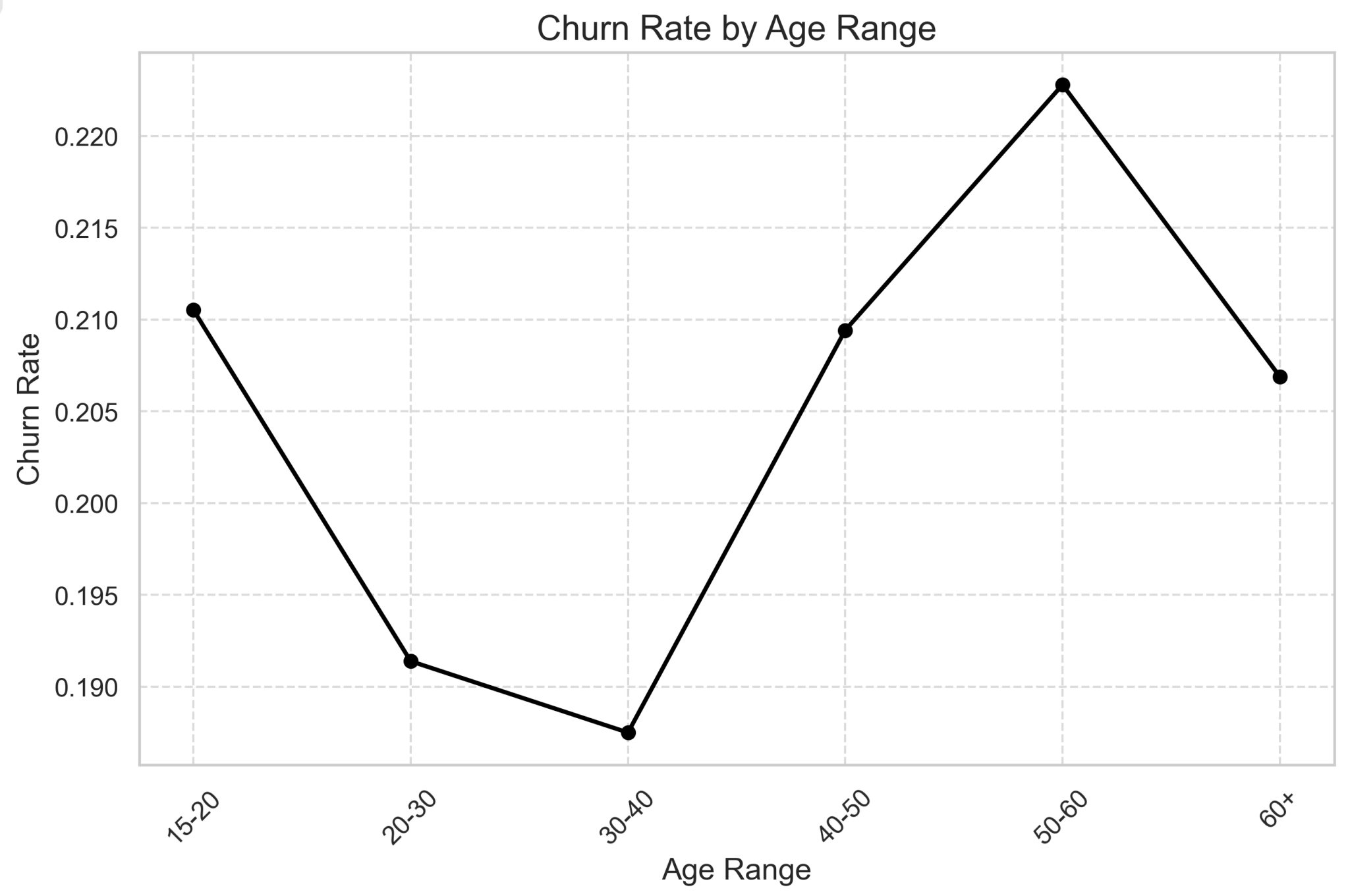
✳**Z-score** transformation and one-hot encoding used for customer demographics including age, gender, marital status, and incomeLevel, ensuring fair feature scaling.

✳ A **Markov Chain** is applied to capture customer patterns of interactions data, including feedbacks, inquiries and compliants.

✳**Proper binning separations** are applied to capture the transaction differences between churn and retained customers.

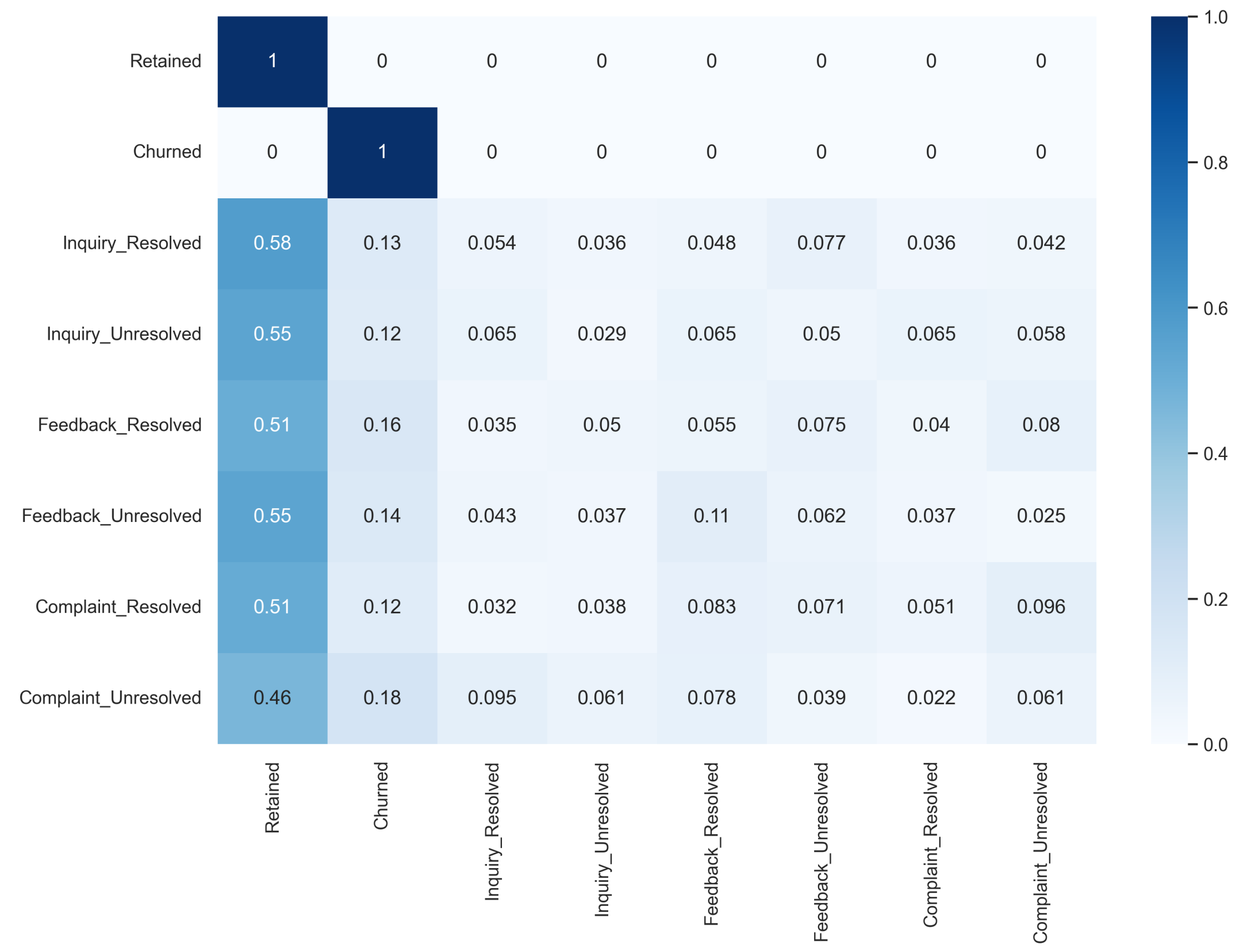
✳A **Student t-test** is used to compare the mean values of continuous features in login activity between churned and retained customer, and the χ^2 **test** evaluates the association between categorical features and customer churn.

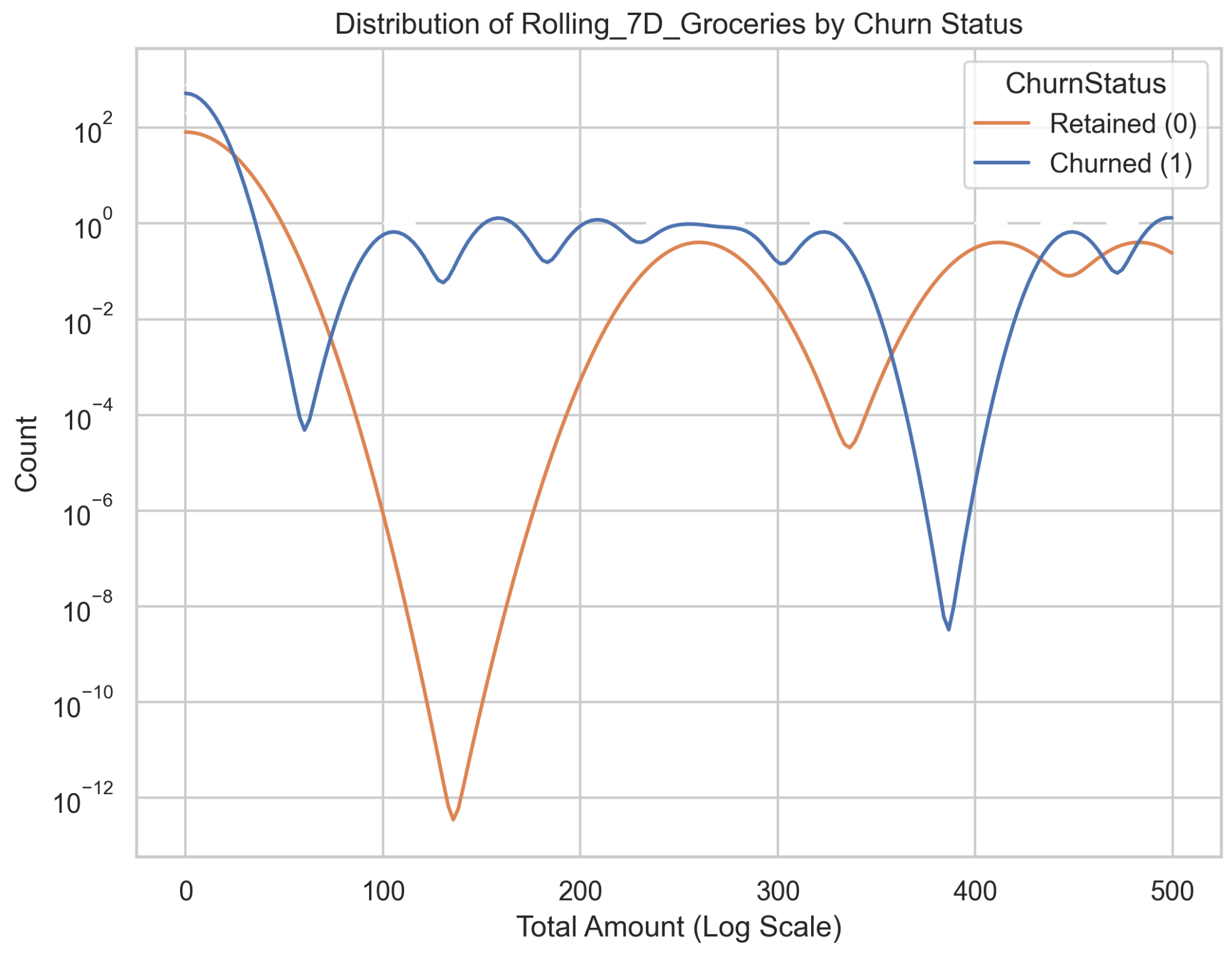
Findings



The age range distribution indicates that the **20-30 and 30-40** age groups have relatively low churn rates compared to other groups. This trend may be attributed to their stable income levels, which contribute to higher customer retention

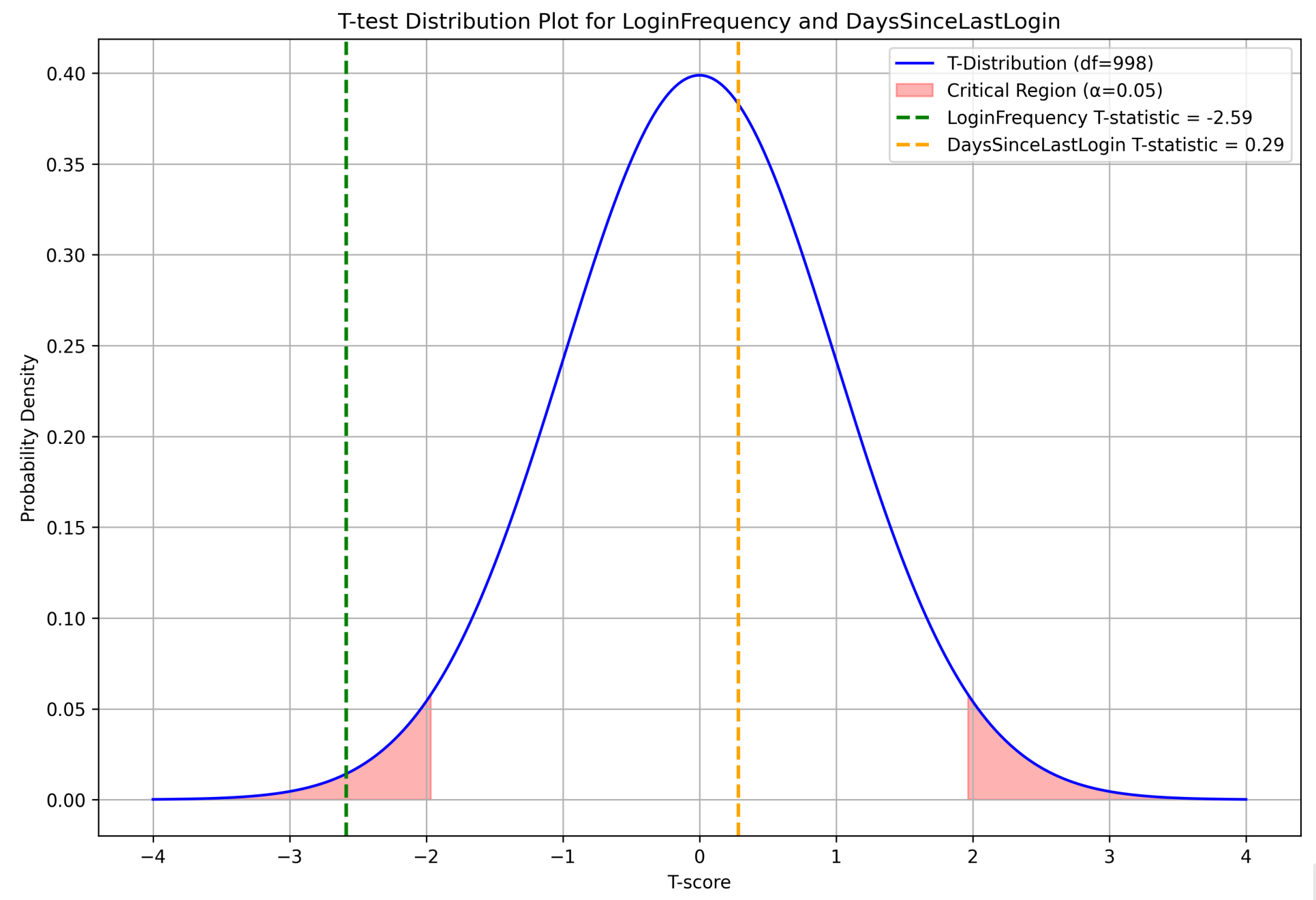
The absorbing probability can be calculated from the transition matrix. '**Complaint Unresolved**' has the highest churn probability (25.52%), suggesting that unresolved complaints are a significant risk factor for customer attrition.





Transaction distribution of **groceries** illustrates the differences in short-time (7 days) purchasing behaviour between retained and churned customers. Proper bining can be used to emphasise the characteristics.

The t-test statistic for **login frequency** falls within the critical region (α = 0.05), leading to the rejection of the null hypothesis that the mean values between churned and retained customers are the same. However, Days Since Last Login does not show a significant difference between churned and retained users.

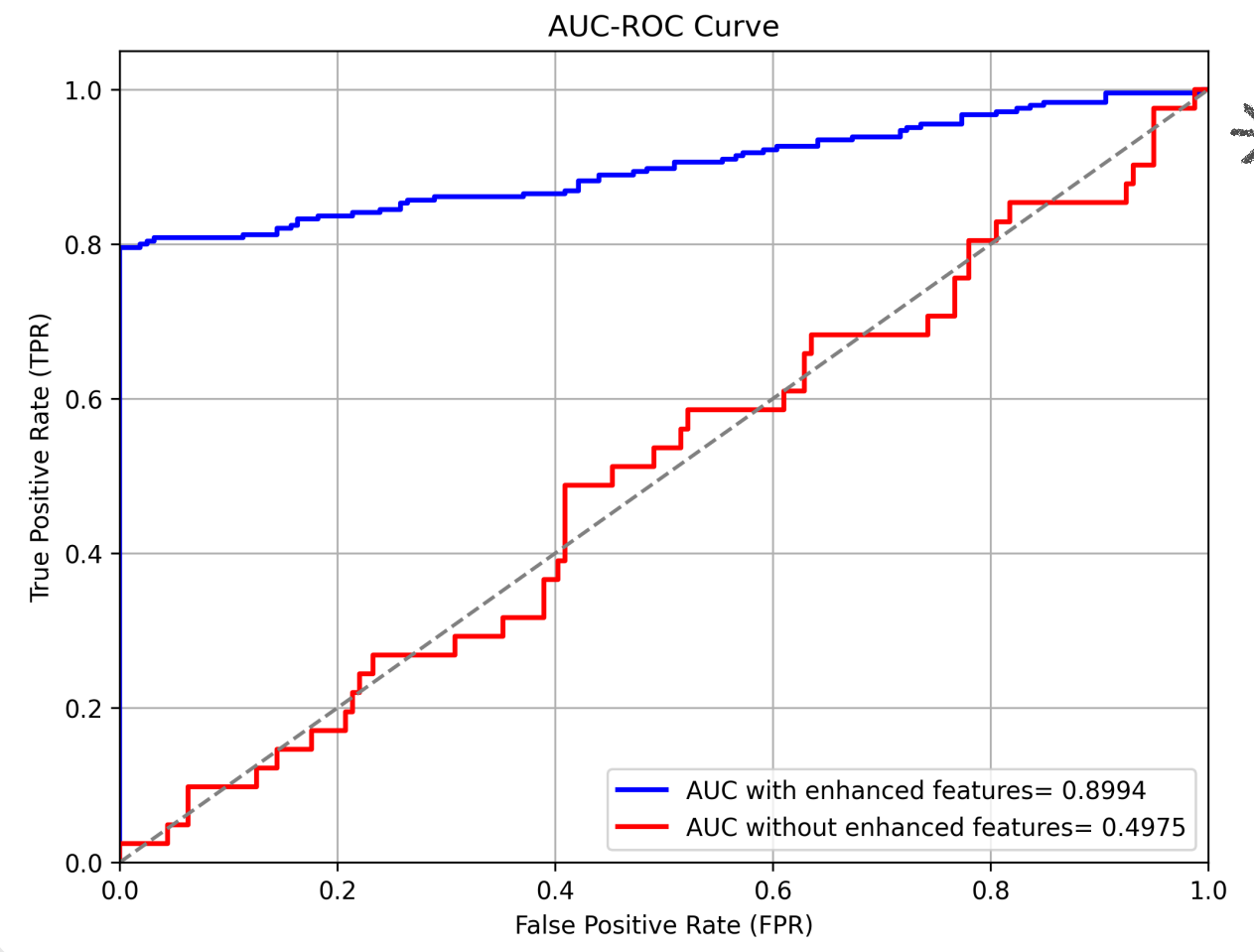


Results

Two XGBoost classifiers are designed to evaluate the impact of statistical feature engineering, comparing performance with and without feature enhancements.

Metrics	Churned	Retained
Baseline Model Precision	0.27	0.80
Baseline Recall Rate	0.10	0.93
Feature Enhanced Model Precision	0.94	0.76
Feature Enhanced Model Recall	0.81	0.92

While the accuracy and recall for the retained class remained mostly unchanged, the **accuracy and recall for the churned class significantly improved**.



✳The **AUC increased from 0.5 to approximately 0.8**, indicating that the feature engineering had a significant impact on improving the machine learning model.

Summary

✳Statistical feature engineering techniques, such as Z-score normalization, Markov Chains, binning separation, and hypothesis testing, can effectively improve the model's performance in terms of customer churn precision and recall rates.

✳ From statistical analysis, factors like unresolved complaints, recent spending behavior, and online engagement, which play crucial roles in churn, can be quickly identified. The findings can also help businesses develop targeted retention strategies.