# Customer Churn Prediction

# and Recommendation System

Emmy Son

# Introduction

# Customer churn is a critical metric for subscription-based services, representing the percentage of customers who stop using the service over a given period. For a streaming service, predicting which customers are likely to churn allows proactive engagement to retain users and reduce revenue loss. The dataset includes demographic, behavioral, and subscription information for customers, such as gender, age, income, months subscribed, plan type, mean hours watched, competitor subscriptions, genre preferences, and historical account behavior (e.g., cancelled or downgraded). The target variable is churn, a binary indicator of whether a customer has left the service. If successful, predictive modeling can identify high-risk customers early, allowing targeted retention campaigns, and the recommendation system created can provide personalized recommendations, and strategic marketing efforts that improve customer lifetime value and satisfaction.

# Methods

# To better understand the factors influencing customer churn, we first explored the available data. This included demographic information (such as gender and age), financial indicators (income), subscription details (plan type, months subscribed, bundle status), content engagement metrics (mean hours watched, longest session, favorite genres), and account history (previous cancellations or downgrades, number of profiles, Kids profile). Initial exploration included inspecting summary statistics and checking for missing values, which helped us identify any potential data issues and understand patterns in customer behavior.

# Prior to modeling, we performed standard data cleaning by removing rows with missing values in key variables like age, income, and past cancellations, and then reset the index. Categorical variables, including gender, plan type, and content genres, were converted into multiple binary variables through one-hot encoding to ensure the models could interpret them effectively. Continuous variables, such as age, income, months subscribed, and hours watched, were standardized using z-scores to normalize differences in scale. The dataset was then split into training and testing subsets (90/10 split) to ensure that model evaluation would reflect performance on new, unseen customers.

# We developed two predictive models. The first was logistic regression, which provides a straightforward way to estimate the probability that a customer will churn based on individual features. This model is highly interpretable, allowing us to understand which factors—such as low engagement, previous cancellations, or certain subscription plans—contribute most strongly to churn risk.

# The second was a gradient boosting classifier, which builds an ensemble of decision trees to capture more complex patterns and interactions among features. This model can uncover subtle relationships, for example: how a combination of content engagement, subscription type, and account history might jointly influence churn. We evaluated both models using standard classification metrics (Accuracy, Recall, Precision, and ROC AUC) to assess their predictive performance. Additionally, calibration curves were used to check how well the predicted probabilities aligned with actual churn outcomes. These insights can help the streaming service target high-risk customers with retention strategies and personalized recommendations.

For the recommendation system, the L2 logistic regression model was used to predict churn probabilities for new customers. The top 200 high-risk customers were then analyzed using a Nearest Neighbors model on a ‘favorites’ dataset, enabling identification of the 10 most similar users to each high-risk customer for targeted content recommendations.

# Results

**Model Performance**

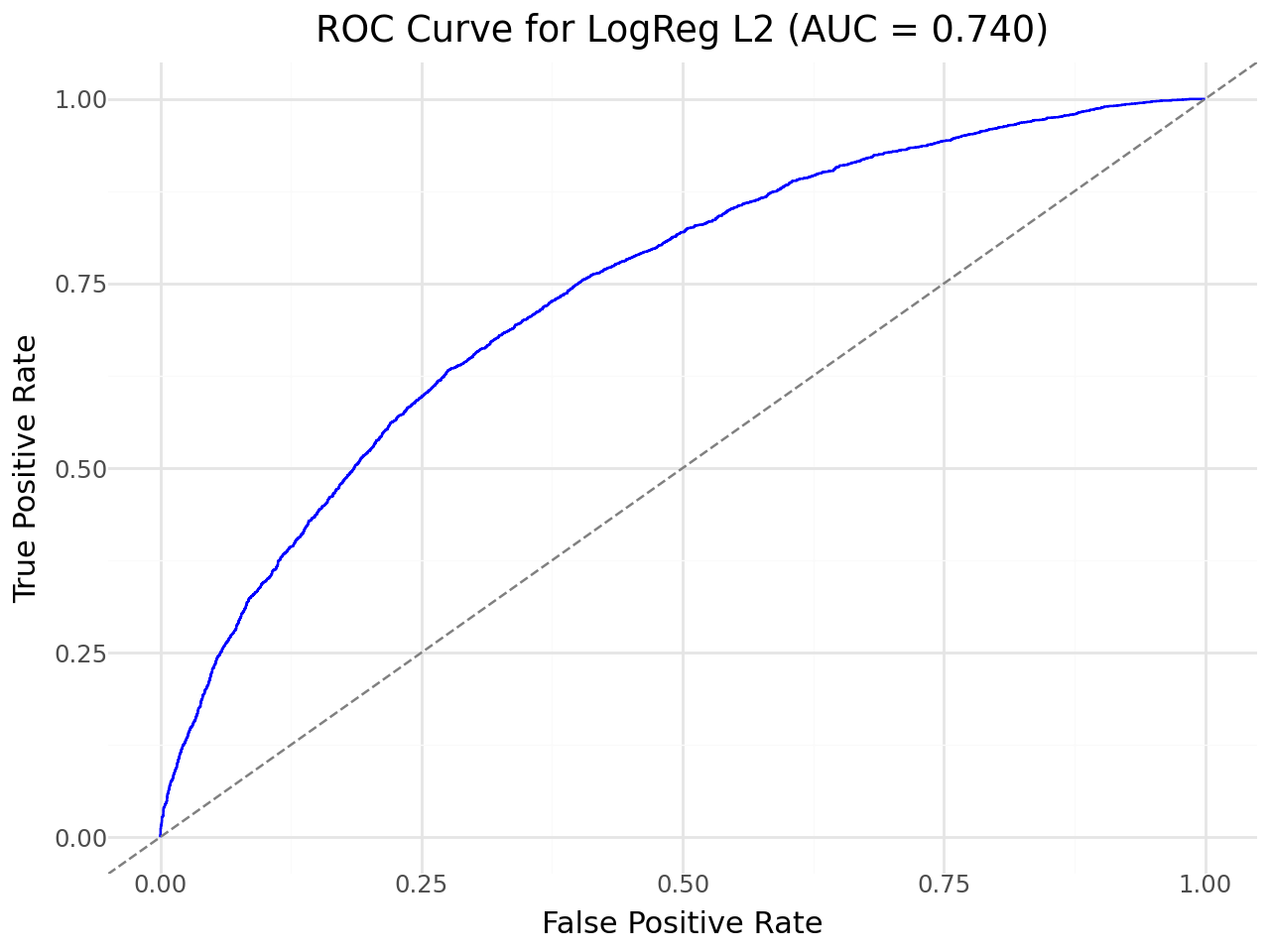
Both logistic regression (L2) and gradient boosting performed similarly, with training and test accuracies around 74%, indicating good generalization and no strong overfitting. Recall was low (~27%), meaning some churners are missed, while precision was moderate (~62%), and ROC AUC was ~0.74 for both models. Calibration curves show predicted probabilities align reasonably with actual outcomes, making probability estimates reliable for identifying high-risk customers. I recommend logistic regression for production due to its simplicity, interpretability, speed, and lower resource requirements. Gradient boosting offers slightly higher precision but is slower and harder to interpret. Using these predictions, the recommendation system identified the top 200 most at-risk customers. For each high-risk customer, their 10 nearest neighbors were found based on age, income, and mean hours watched, producing a dataset ready for personalized content suggestions. This system ensures that retention efforts can be prioritized for the most vulnerable customers while increasing engagement with tailored recommendations.

**CEO Recommendations:** Use logistic regression to identify high-risk customers for retention campaigns. Combine churn probabilities with targeted initiatives like personalized content or subscription offers. The recommendation system can suggest films enjoyed by similar users, increasing engagement and reducing churn risk.

**Logistic Regression Results Table:**

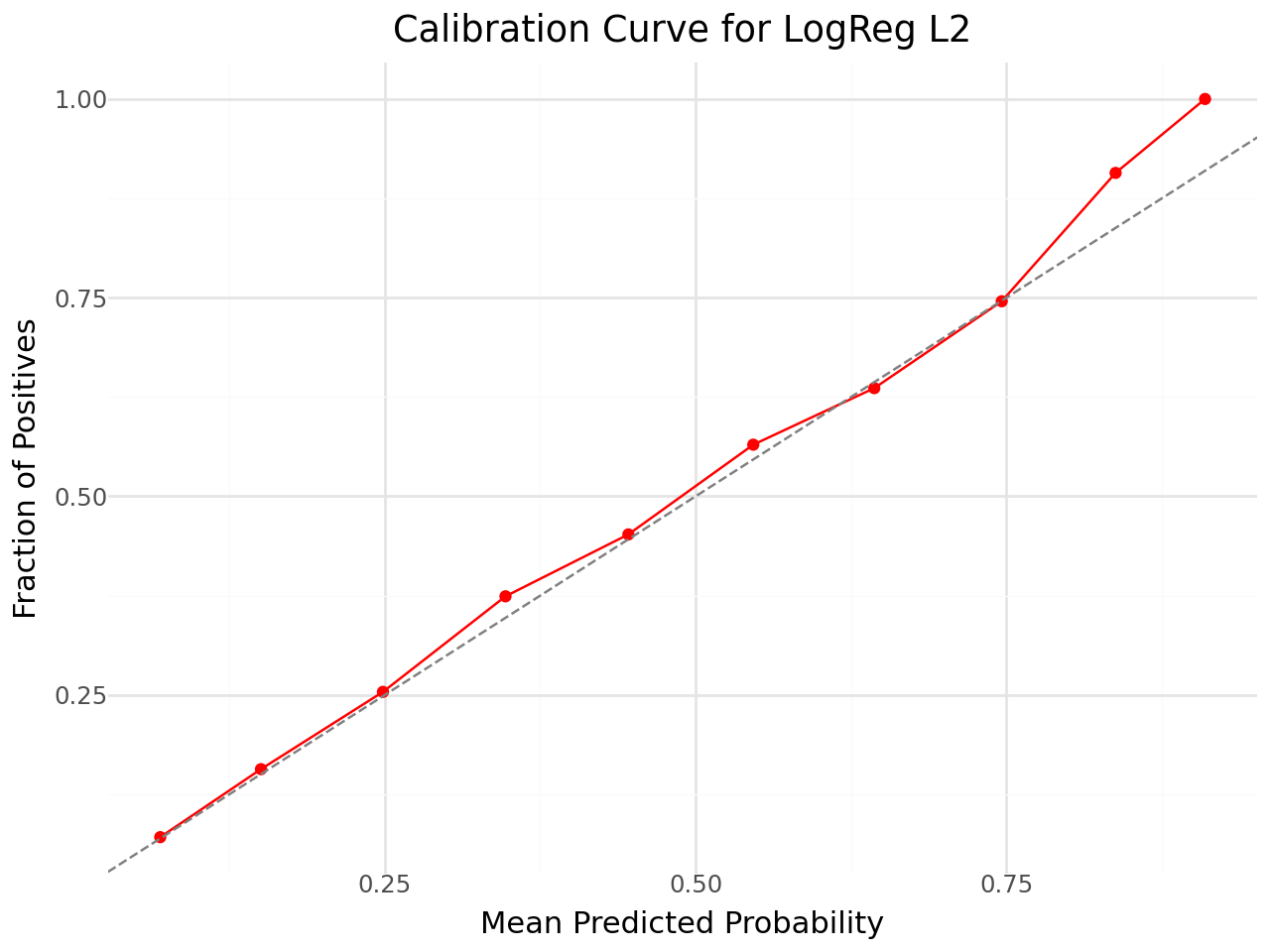
| **Metric** | **Train** | **Test** | **Interpretation** |
| --- | --- | --- | --- |
| Accuracy | 0.7416 | 0.7369 | Correctly classifies ~74% of customers |
| Recall | 0.2744 | 0.2821 | Low recall; many churners missed |
| Precision | 0.6052 | 0.6203 | Moderate precision; predictions of churn are ~62% accurate |
| ROC AUC | 0.7354 | 0.7400 | Good discrimination ability between churners and non-churners |

**Log Regression with L2 Penalty Model Performance Visuals:**



**Figure 1:** ROC Curve for Logistic Regression with L2 Penality

The curve illustrates the model’s ability to distinguish between customers who churn and those who do not, with an AUC of 0.740 indicating moderately strong discrimination.



**Figure 2:** Calibration Curve for Logistic Regression with L2 Penalty

The model’s predicted probabilities align closely with the observed outcomes, indicating good calibration overall, though it slightly overestimates churn for customers with high predicted probabilities

**Gradient Boosting Results Table:**

| **Metric** | **Train** | **Test** | **Interpretation** |
| --- | --- | --- | --- |
| Accuracy | 0.7440 | 0.7360 | Correctly classifies ~74% of customers |
| Recall | 0.2655 | 0.2687 | Low recall; many churners missed |
| Precision | 0.6217 | 0.6238 | Moderate precision; predictions of churn are ~62% accurate |
| ROC AUC | 0.7400 | 0.7389 | Good discrimination ability, similar to logistic regression |

# Discussion/Reflection

Through this analysis, I gained a deeper understanding of how customer demographics, subscription behavior, and viewing patterns can influence churn. I also learned the importance of model selection in respect to comparing logistic regression and gradient boosting in terms of trade-offs between interpretability, complexity, and performance. If I were to perform this analysis again, I would explore additional features, such as engagement over time or interaction effects between genres and plan type, to improve recall and capture more churners. I would also consider hyperparameter tuning for gradient boosting and experiment with ensemble methods to balance precision and recall more effectively. In addition, the recommendation system added practical value by using churn predictions to select the top 200 high-risk customers and identify similar users for content suggestions. In the future, I would explore more sophisticated similarity metrics or additional behavioral features to improve the quality of recommendations and further reduce churn.