Subscription Churn Rate Prediction for OTT Platforms

Kashish Thakur, Sourab Saklecha, Swetha Neha Kutty Sivakumar, Yuti Khamker

Motivation

In the competitive arena of subscription-based music streaming services, predicting user churn accurately is essential for sustaining profitability. Minor fluctuations in churn rates can lead to substantial financial consequences. By understanding and forecasting which users are likely to cancel their subscriptions, businesses can implement targeted strategies to retain customers, enhance engagement, and boost revenue.

Background

KKBOX is Asia's foremost music streaming service, featuring an extensive collection of over 30 million Asia-Pop tracks. They operate a hybrid model that offers both ad-supported free access and premium paid subscriptions. The success of this model depends heavily on retaining their paid subscribers, making it essential to anticipate and mitigate churn effectively. By leveraging machine learning techniques and algorithms, this project aims to enhance the prediction of user churn, allowing KKBOX to refine its strategies for reducing customer attrition, increasing user engagement, and ultimately driving revenue growth.

Literature Review

The study conducted by Hardjono and Isa (2022) addresses the high churn rate of a music streaming provider, where 90% of customers uninstalled the app compared to new installations. To analyze and predict customer churn, researchers implemented data mining techniques and tried comparing machine-learning models: XGBoost and logistic Regression with hyperparameter tuning on the 2020 dataset from Google Analytics. The result reported the highest-performing model was Logistic Regression with Hyperparameter Tuning which predicted customer churn with 94.77% accuracy and a 97.03% F1-score.

The paper "Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform" by Ahmad et al. (2019) built a churn prediction model for SyriaTel, focusing on using social network analysis (SNA) and feature engineering in a big data environment. Several data sources, including call detail logs, customer records, and mobile IMEI information were processed through the Hortonworks Data Platform having HDFS for storage and Spark for processing. The important elements in the project were: statistical aggregation and degree centrality and cosine similarity as their SNA metrics. The authors employed four machine learning models: Decision Tree, Random Forest, GBM, and XGBoost incorporating cross-validation and hyperparameter tuning. Predictive accuracy was considerably increased by the incorporation of SNA characteristics. When the strategy was implemented, SyriaTel was able to retain 47% of high-risk churn, demonstrating a major business benefit.

According to Chen's (2019) research, customer churn prediction is an important factor in the long-term viability of subscription businesses. Based on a sizable customer dataset, a parallel artificial neural network is constructed in the study to produce a highly accurate customer churn model. It addresses the scalability and efficacy of the suggested methodology and has produced noteworthy accuracy results.

Methodology

Data Collection

The dataset for this project was gathered via Kaggle. With elements like user demographics, membership durations, payment methods, and activity logs, the dataset offers a thorough perspective of user interactions and subscription details on KKBOX's platform. The original dataset comprised several files including user demographics, transaction records, and user log data. These raw, dirty files needed extensive cleaning and preparation since they

contained inconsistent and partial data entries. Moreover, amalgamating these several files was required to create an extensive dataset fit for the study. The dataset can be found here.

Exploratory Data Analysis

A thorough examination of user behavior and subscription trends is needed for initial understanding and cleaning. The churn rate was particularly high among younger users, peaking at over 35% for those around 20 years old, according to our data, which revealed numerous important results. Additionally, we discovered that customers who had auto-renew activated had a far lower churn rate—less than 20% vs almost 80% for those who did not. Furthermore, the data on city distribution revealed significant differences in turnover rates between cities, varying from 5% to almost 20%. The examination of payment methods showed that churn rates for some of the methods might reach 60%, indicating that certain transactional methods may not be as effective for long-term retention. These measurements are essential for creating focused tactics that cater to the distinct qualities of different user categories.

Data Cleaning

The 'train' and 'members' datasets were first combined, and we improved them by deleting any duplicate columns and records with missing values. Additionally, memory reduction was done. After that, we added computed metrics like payment discrepancies and user engagement dates and statistics to the comprehensive transaction and user log data that had been collected using the index ('msno'). By following these procedures, a clean, complete dataset that was ideal for building reliable churn prediction models was produced.

Feature Engineering and Transformation

We enhanced our dataset through meticulous feature engineering. We generated additional temporal variables like the age of the account at a given cutoff date and the number of days until membership expiration. Next, we used one-hot encoding to encode categorical variables, such as city, registration method, and payment method, so that machine learning algorithms could use them. We used the StandardScaler to scale numerical features, omitting the target variable "is_churn," To resolve the class imbalance, we finally employed both upsampling and undersampling strategies, which improved the predictive models' resilience against overfitting and underrepresentation of minority classes. Finally, the data was prepped for modeling by performing a 70-10-10 train-validation-test split.
A detailed workflow diagram can be found here

Model Details and Training

Support Vector Machine Model: According to Zhang (2012) SVM is a statistical learning theory, which is used for Class Classification problems. The current project utilizes this approach to classify whether users will churn or not. The SVC classifier is created using a Linear Kernel function, with a regularization parameter, the value of which is identified using experimentation, and maximum iteration is set to 100,000.

Random Forest Model: It is a great option for churn prediction as it can handle high-dimensional datasets, reduce overfitting through ensemble learning, and is robust against imbalanced data. It offers interpretability and efficiently manages missing values through feature priority scores. A Random Forest model with tuned hyperparameters is used to train and evaluate comprehensible predictions for identifying potential churners.

Logistic Regression Model: The statistical model performs best for classification tasks to determine whether a user will churn or not. The optimal parameters are determined by the model in order to maximize the likelihood of the data. Because of this, logistic regression is a

strong machine-learning technique for determining how input factors and binary outcomes are related to prediction. Logistic regression is trained and regularizations are applied to improve the performance.

XGBoost Ensemble Model: The XG Boost model uses a gradient boosting process to create several decision trees in order, with each tree resolving the errors of the one before. As a result, a strong ensemble model is produced that is capable of capturing intricate relationships in the data. The XGBoost model is trained with 500 iterations with ridge regularization to improve model performance.

Experiment and Results

The number of values in each class of the 'is_churn' column is: For class 0 there are 679119 records, and for class 1 there are 46603 records. There is a huge imbalance between the two classes of the target column. To evaluate the effectiveness of different models for churn prediction, a few experiments were conducted addressing the challenge of the imbalanced dataset.

Support Vector Machine Model

The model is initially trained on imbalanced data, with a Regularization parameter of 0.1, and a Linear kernel function. The value of the Regularization parameter is identified by analyzing the variation in accuracies for different values of the Regularization parameter, which were 0.1,0.01,1 etc. The most optimal regularization parameter was found to be 0.1. The model trained on imbalanced data achieves an accuracy of 71.6%, with F1 Score of 62.23% , with Precision and Recall having a value of 91.76% and 47.08%. The results of the model are significantly improved when the model is trained on data balanced via undersampling. It achieves the accuracy of 88.9%, with F1-Score of 89.02%, with Precision and Recall having the value of 87.54% and 90.54%. Hence we can conclude that the SVM model performs better with balanced data.

Random Forest Classifier Model

Because of inherent robustness to class imbalance, the model was trained on imbalance data by employing 200 trees in the Random Forest classifier to predict subscriber churn, with a maximum depth of 10 for each tree to prevent it from overfitting. A node has a minimum of 50 split samples. Furthermore, a minimum of 20 samples are needed for every leaf node, which smoothes out predictions. The trained model is then used for churn prediction on test data. The model was evaluated on the mentioned metrics achieving an accuracy of 89.4%, F1-score of 88.1% and ROC value of 0.89. Additionally, balanced evaluation across both classes was guaranteed via stratified 10-fold cross-validation. The second experiment was to train the model on balanced data where undersampling techniques were incorporated wherein the majority class was undersampled. This was achieved by randomly sampling the majority class without replacement and then a new balanced data was created by combining the undersampled majority class with the minority class. By mitigating the bias towards the majority class, the models achieved an accuracy of 97% , F1-Score of 96.8% and ROC value of 0.99.

Logistic Regression Model

The logistic regression model was initially trained on imbalanced data which achieved an accuracy of 83%. The model was retrained using balanced data using an undersampling strategy to balance the target class in order to enhance its performance. The model obtained a 71% accuracy rate by employing the undersampling strategy. Although there has been improvement in the log loss when compared to imbalanced data, the random guessing approach still outperforms it. The following experiment used the Synthetic Minority

Over-sampling (SMOTE) Technique, which oversamples occurrences of minority classes relative to the predominant class, to balance the data. Using this method, the model produced a 90% accuracy.

XGBoost Model

The XGBoost classifier is trained on imbalanced data as well as balanced data with a learning rate of 0.01, determining the step size for updating weights and the number of trees in the ensemble model set to 500, and each tree is to a maximum depth of 3. The objective is set to binary logistic indicating the binary classification and ridge regularization is set to avoid oversampling. The experimental results show both balanced and imbalanced data techniques achieved the same 88% accuracy.

Results

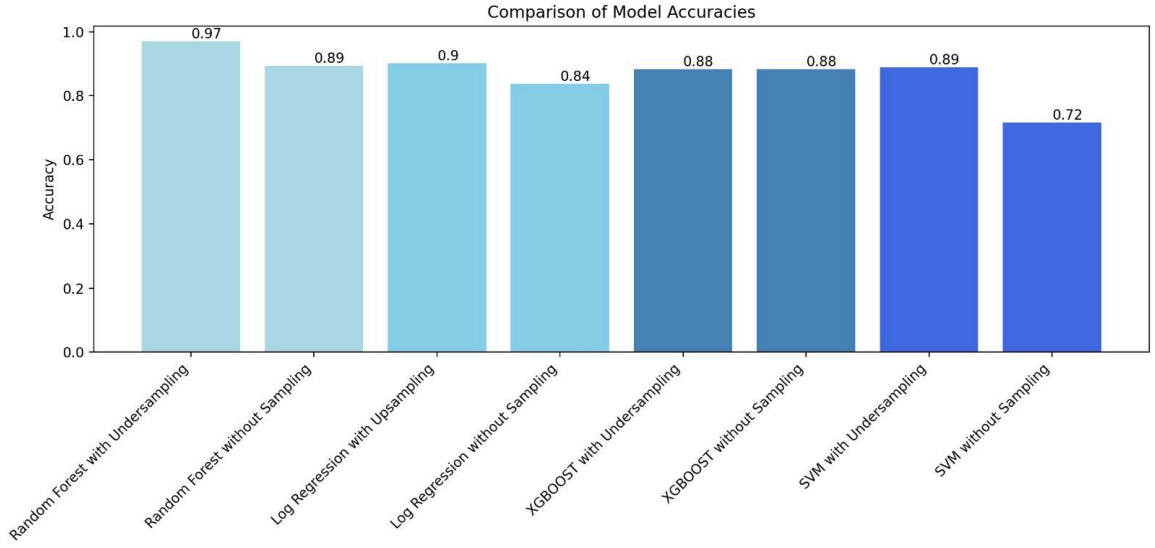
Figure 1

Model Comparison Results using Comparison Table

	Model	Accuracy	Precision	Recall	F1 Score
0	Random Forest with Undersampling	0.970	0.969819	0.969819	0.969819
1	Random Forest without Sampling	0.894	0.987531	0.796781	0.881960
2	Log Regression with Upsampling	0.901	0.911157	0.887324	0.899083
3	Log Regression without Sampling	0.837	0.977143	0.688129	0.807556
4	XGBOOST with Undersampling	0.884	0.989717	0.774648	0.869074
5	XGBOOST without Sampling	0.884	0.989717	0.774648	0.869074
6	SVM with Undersampling	0.889	0.875486	0.905433	0.890208
7	SVM without Sampling	0.716	0.917647	0.470825	0.622340

Figure 2

Comparing the Model Accuracies using Bar Plot



The results obtained after evaluating the models on the Test set are described in the provided figures. Figure 1 provides the model comparison on the basis of different evaluation metrics. Figure 2 shows a bar plot showing the model's comparison based on their accuracies. Followed by Figures 3 and 4 showing the most optimal ROC curve for the four models.

Figure 3

ROC Curve for SVM Model and Random Forest Model

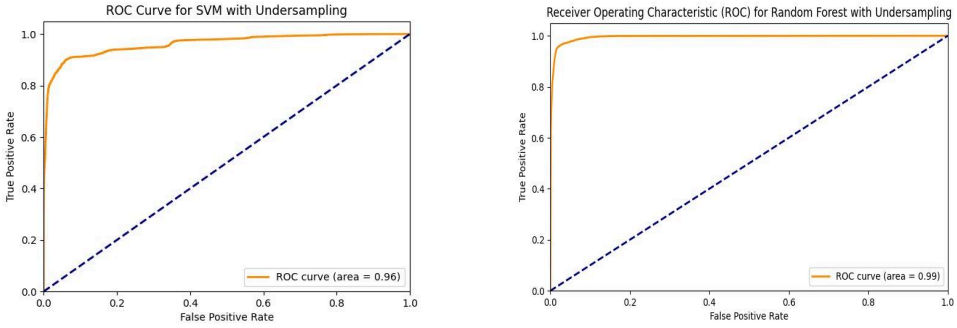
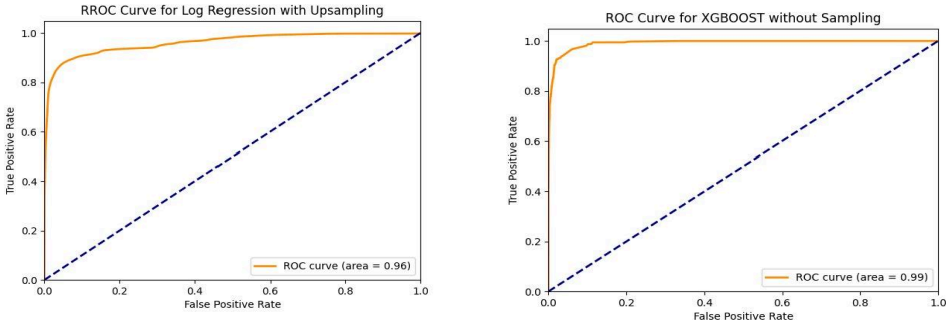


Figure 4

ROC Curve for Logistic Regression Model and XGBoost Model



Discussion & Future Improvement

The KKBox subscription data was used for this research together with both balanced and unbalanced data methodologies to train machine learning models such as SVM, RandomForest, logistic regression, and XGBoost. This experiment's maximum accuracy of 97% was achieved by the RandomForest model using the undersampling strategy. This result emphasizes how tailored data balancing techniques can improve model performance on real-world datasets.

A number of approaches can be taken in the future to increase the churn prediction model's efficacy and applicability.

Expanding to Additional Subscription Platforms: Generalizing the model to accommodate data from different Subscription platforms can increase its adaptability and suitability for a greater variety of use cases. This entails modifying the model to accommodate various content consumption patterns and platform-specific user interactions.

Developing Real-Time Prediction System: As user behavior and market dynamics might alter over time, it is imperative to test the prediction model on more recent data to guarantee its applicability and accuracy. Maintaining the model's predictive strength and dependability will require routinely adding new data to it.

Applying More Advanced Deep Learning Algorithms: Enhancing the churn prediction by utilizing advanced deep learning models. Techniques like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) will help us capture complex patterns for more accurate predictions. This approach promises improved performance and deeper insights into customer behavior.

References

Ahmad, A.K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using machine learning in a big data platform. *J Big Data* 6, 28 (2019). https://doi.org/10.1186/s40537-019-0191-6

Chen, M. (2019). Music Streaming Service Prediction with MapReduce-based Artificial Neural Network. *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 0924–0928. https://doi.org/10.1109/uemcon47517.2019.8993063

Hardjono, C., & Isa, S. M. (2022). Implementation of data mining for churn prediction in music streaming companies using 2020 dataset. *Journal on Education*, *5*(1), 1189–1197. https://doi.org/10.31004/joe.v5i1.740

Zhang, Y. (2012). Support Vector Machine Classification Algorithm and its application. In *Communications in computer and information science* (pp. 179–186). https://doi.org/10.1007/978-3-642-34041-3_27