Churn Prediction for a Music Streaming Service Company

Ahana Roy Choudhury, Chaitali Pawar, Nitin Sahai, Saurabh Suman, Rahul Gaadhe, Matthew A. Lanham
Purdue University Krannert School of Management
aroycho@purdue.edu;cpawar@purdue.edu;nsahai@purdue.edu;ssuman@purdue.edu;rgaadhe@purdue.edu;lanhamm@purdue.edu

Abstract

For a subscription business, accurately predicting whether a subscriber will churn after their subscription expires is critical to profit and long-term success. The dataset and business problem statement is provided and defined through Kaggle. We formulated this as a predictive analytics problem with target output (is_churn) - a binary categorical variable with values of 0 and 1. We built four models and the algorithm that performed with the highest prediction accuracy rate was chosen.

Introduction

Churn is a very critical problem in business. Companies spend millions of dollars to acquire a customer. So, when the customer churns or moves to any other competitor company, or discontinues the service of the company, it is a huge financial loss for the company. Business organizations want to know beforehand which customers are the most prone to churning. This information helps an organization to plan its resources better, for example, creating a customized marketing plans for a specific customer so as to prevent that particular customer from churning.

Research question:

How well can one predict a churner and what are the drivers of churning?

Literature Review

The customer churn problem has been widely studied. Most of these studies follow the CRISP-DM methodology. Studies show that nonlinear ANN performs better than ANN, which performs better than decision trees, which in turn works better than logistic regression. Techniques based on statistical methods like Linear Regression, Logistic Regression, K-Means and Naïve Bayes also works well. Churn being a categorical outcome, Logistic Regression works fine. But often advanced techniques like decision trees, rule-based learning and ANN are used in conjunction with a simple logistic regression. Naïve Bayes is a supervised learning module which makes prediction based on Bayes theorem. There are a few papers we read which have also established that decision tree approach may outperform the ANN approach. In some cases, ANN out performs the decision tree approach. This can happen because of the size of datasets used and different feature selection methods applied.

SL No	Paper	Modeling Techniques Used	Metric of Performance
1	Customer event history for churn prediction: How long is long enough?	Logistic Regression, CART, C5.0, CHAID	Classification Accuracy, AUC, ROC, Top Decile and Overall Accuracy
2	Turning telecommunications call details to churn prediction: a Data Mining approach	Binomial logistic regression	2-Log-Likelihood
3	Customer churn prediction using improved balanced random forests	Improved balanced random forests (IBRF)	lift curve and top-decile lift
4	Applying data mining to telecom churn management	Decision tree, neural network, K-means cluster	RSquare, MSE hit ratio , capture ratio ,the decile LIFT of 9.53%.
5	Churn prediction in subscription services - An application of SVM while comparing two parameter selection technique	SVM	AUC
6	Unsupervised Feature Learning on Abstract Company Independent Feature Vectors	Neural Network	Accuracy

Table 1. Literature review summary by techniques and performance metric

Our study is novel because we compare and contrast all the methods used previously, but also ensemble them together.

Methodology

Figure 2 outlines our study design, starting from data collection, data cleaning, data pre-processing, feature creation and selection, model/approach selection, cross-validation design, and model assessment/performance measures.

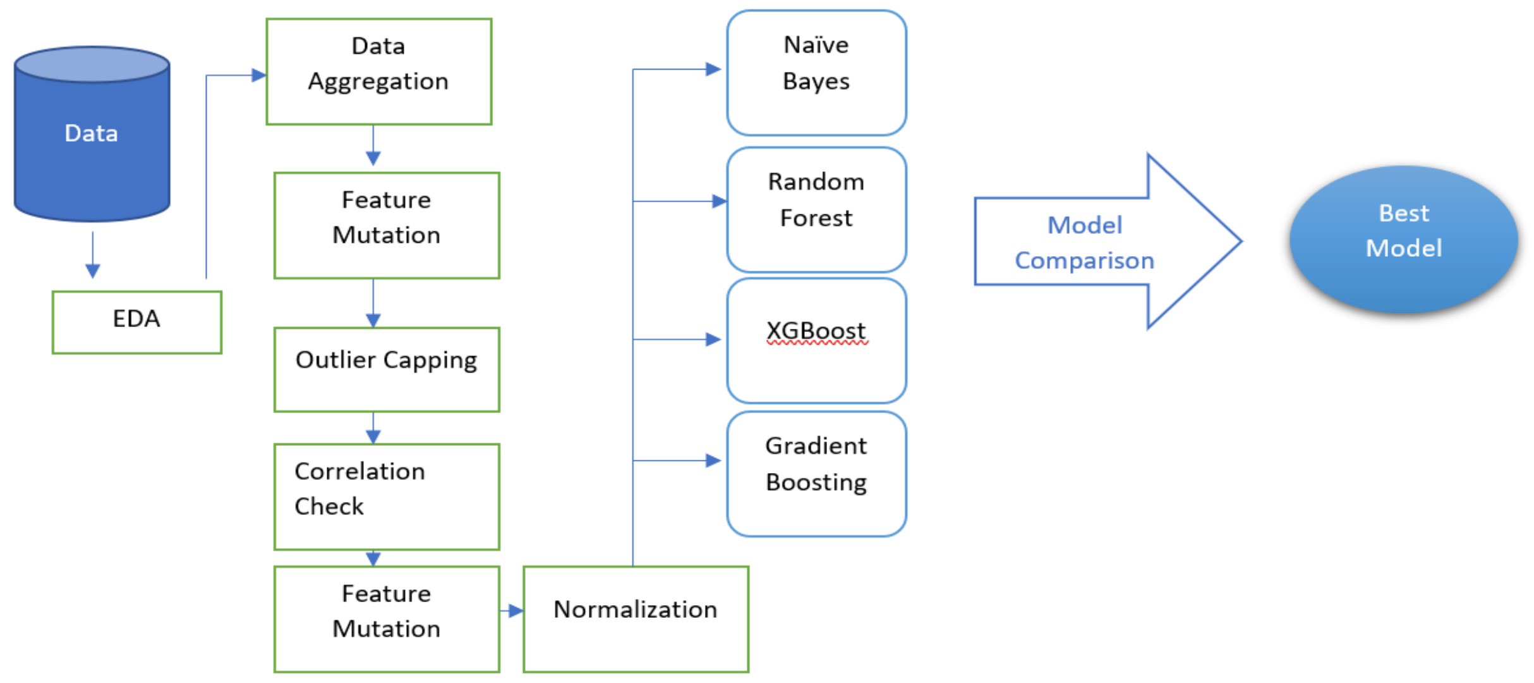


Figure 2. Study Design

Data

The data in this study was obtained from Kaggle.com. There were 4 different files which had information about various aspects of users and their listening pattern. These files are **transactions**: this dataset has information about user transactions; **user logs**: daily user log of users; **members** has information about the users; The **train** set contains the user ids and whether they have churned.

Feature Selection

Our model calculates the feature importance. The following graph shows the top 10 features in order of importance.

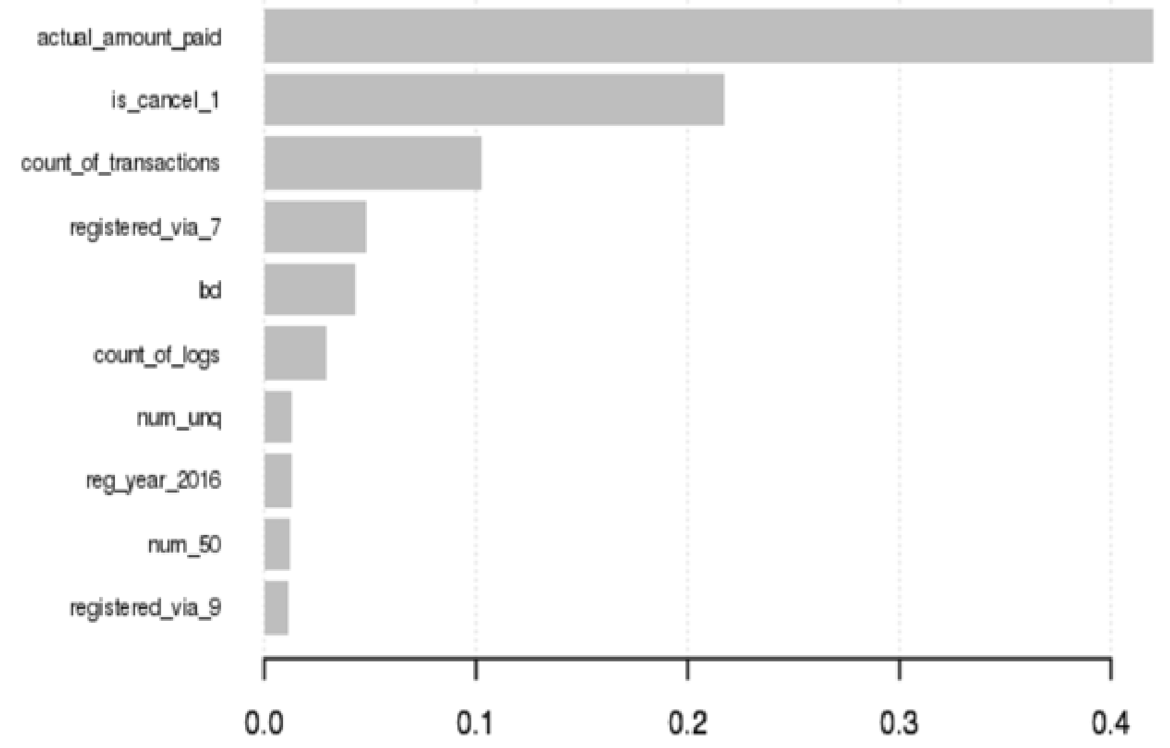


Figure 3. Importance of the features in our model

Model Design

There were 4 different model trained and evaluated. These were Random Forest , XGBoost , Gradient Boosting and Naïve Bayes. The validation method used for tuning the models were different for different approaches : Validation Set approach was used with 60% Train ,20% cross-validation and 20% test split was used for XGBoost 5-fold cross validation was used for Random Forest 70% train ,30% test split was used for Naïve Bayes Gradient Boosting used 3-fold cross validation The data was partitioned using stratified sampling method of createDataPartition() in **caret** library. The business performance measures taken into account are AUC, Balanced Accuracy, and F-1 Score.

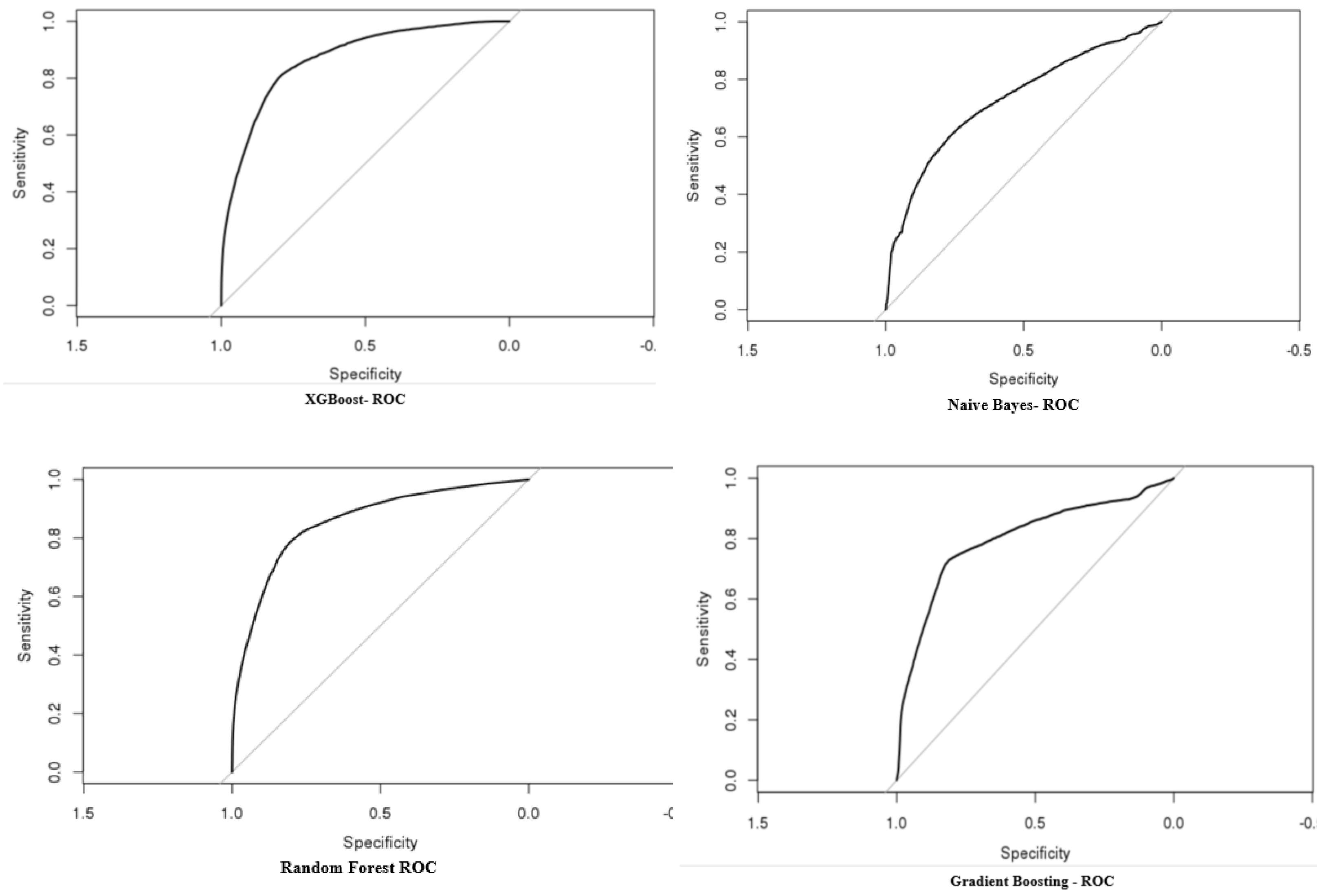
Results

The performance metric of our four models are summarized in the table below.

	AUC	F1	Balanced Accuracy
Naïve Bayes	0.7297	0.9351	0.6053
Random Forest	0.8563	0.9489	0.5876
Gradient Boosting	0.7912	0.9582	0.5901
XGBoost	0.8793	0.9590	0.6008

Figure 4. Model Evaluation

The best performing model was XGBoost with almost similar test and train results. Also it leads in better AUC and F1 compared to other three models.


XGBoost- ROC

Naive Bayes- ROC

Random Forest ROC

Gradient Boosting - ROC

Figure 5. ROC curves for our models

Conclusions

Thus, we were able to predict the churn rate of the customers after their subscription was expired which directly affects the profitability of KKBox music streaming service. We also found that the amount paid and if they had cancelled in the past were top drivers if they would churn in the future. Some actions that the company could consider taking are offering a lower price or promotion to keep price sensitive shoppers. Also, connecting with those that have canceled previously could also reduce the chance of future churning.

Acknowledgements

We thank Professor Matthew Lanham for guidance on this project.