CHURN RATE PREDICTION

KKbox

CAPSTONE REPORT

July 4, 2022

BY: DANIELA MAGIRICU

Table of Contents

KKBox Company Intro

KKBox is a Taiwanese company offering a music streaming service to millions of people across Asia. The company is the leading provider of Asian Pop music, with over 30 million tracks. They offer an unlimited version of their service to millions of people supported by advertising and paid subscriptions.

Business Problem

KKBox would like to forecast the likelihood of customers discontinuing their subscriptions in the future. Their business model is dependent on accurately predicted the churn of their users.

Business Question:

How can KKBox improve customer retention and minimize revenue loss with machine learning?

Business Value

Having the ability to accurately predict future churn rate helps the business gain a better understanding of future expected revenue. By using churn prediction to forecast the potential churn rate of customers, it allows the business to target individual customers to prevent them from discontinuing their subscription. Since the cost of aquiring new customers is higher than keeping existing ones, the business should focus their efforts towards retaining existing customers. Churn prediction is essential for the business to understand what preventative steps are necessary to ensure lost revenue is minimized. Churn prediction is also one of the key components in determining the lifetime value of customers.

Data Source

The dataset is provided by KKBox and was last updated March 2017. It contains 5 CSV files that were used for the purposes of this project and can be accessed here:
https://www.kaggle.com/competitions/kkbox-churn-prediction-challenge/data

Data Cleaning and Modelling Process

The process consists of the following steps: exploration of the dataset with frequency distributions, followed by pre-processing the data, and cleaning the data. This includes determining the shape of the datasets, handling missing values, deleting duplicate rows, and merging the following datasets together into one final dataset:

Members

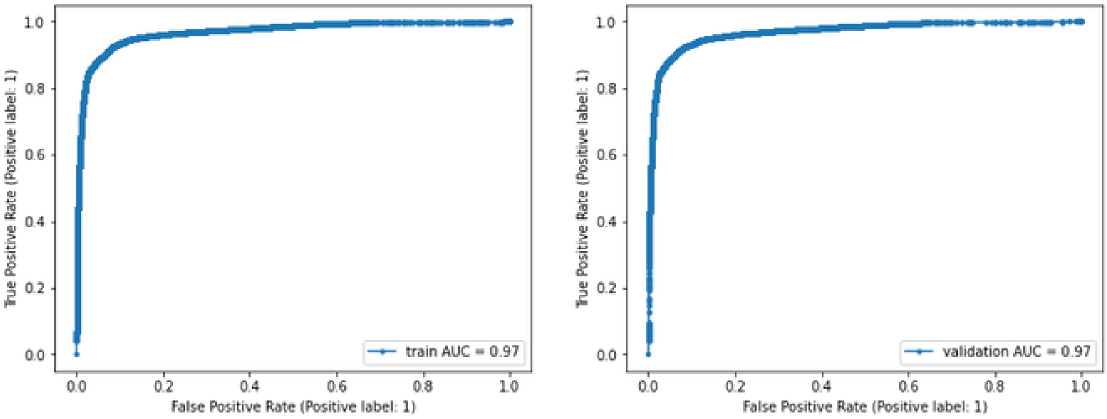
Transactions

User Logs

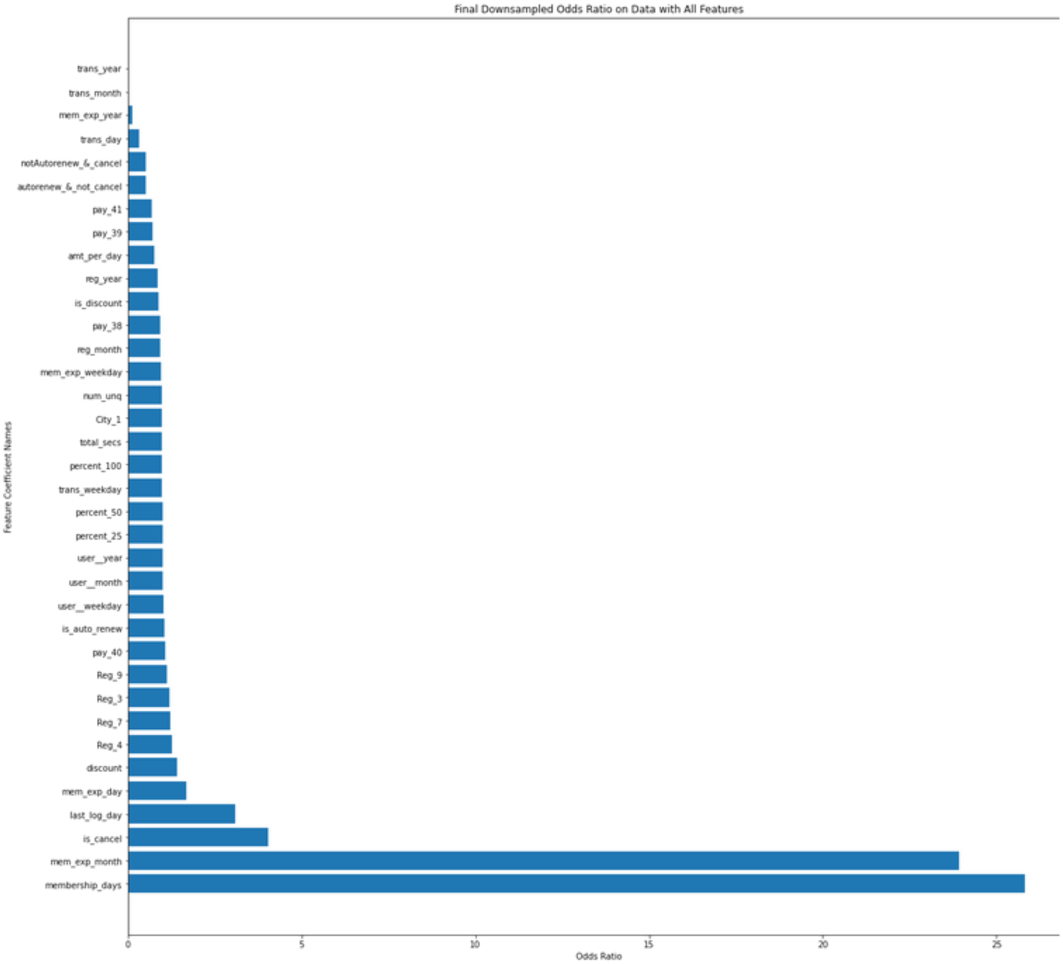
After obtaining the merged dataset new features arecreated. Further cleaning and statistical testing is performed including chi-squared test and correlation heatmaps. A separate dataset is obtained via feature selection where multicollinearity is removed for logistic modelling. Since the dataset has class imbalance in the target variable, churn, upsampling and downsampling techniques are applied. Finally, Logistic Regression Models are scaled, optimized and fit to both the imbalanced and balanced data, and scored on the soft predictions for AUC scores and hard predictions for precision, recall, and f1-scores on the validation set. The same steps are applied to Random Forest Models.

Findings and Visuals

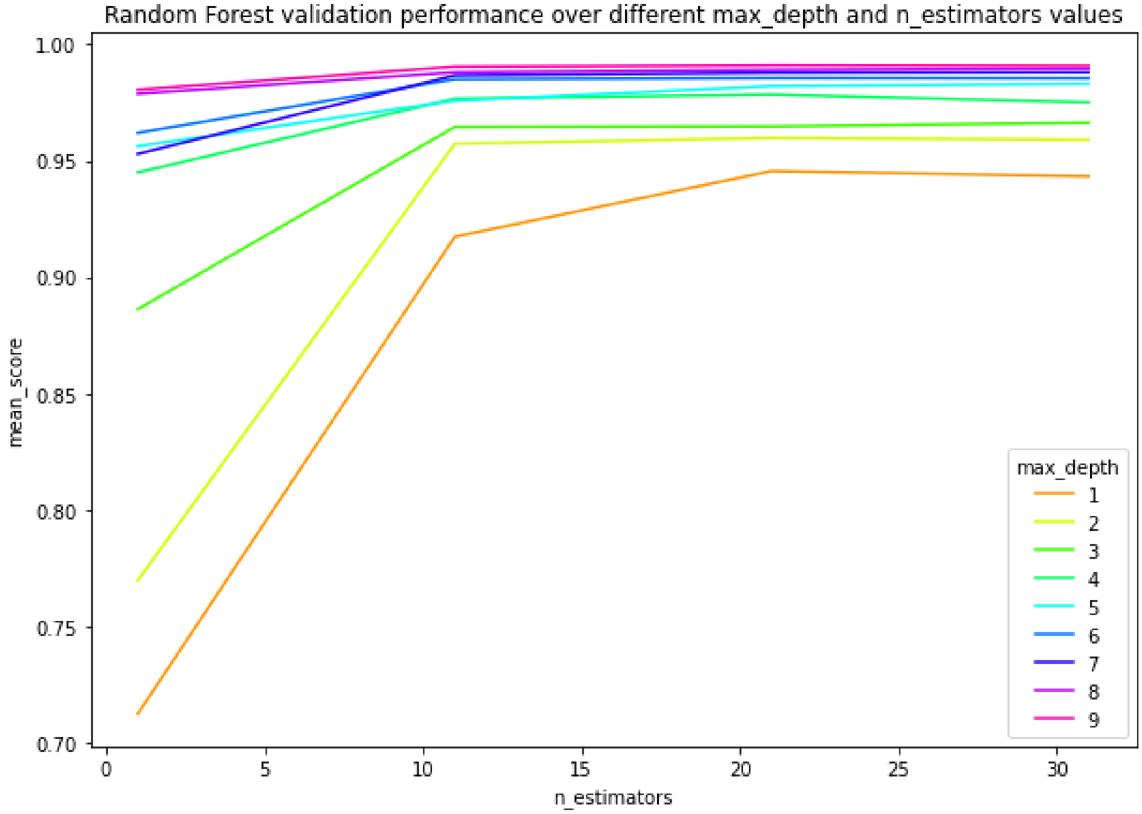
Logistic Regression Model: ROC-AUC Curve



Logistic Regression Model: Odds Ratio



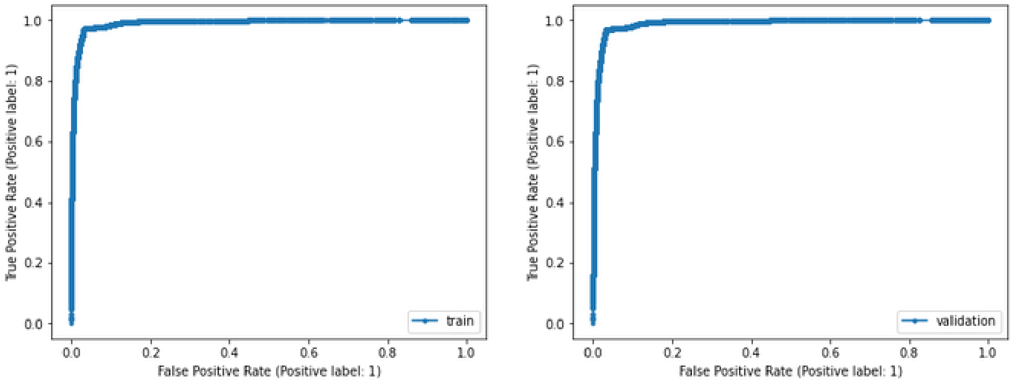
Random Forest Model: Hyperparameter Optimization



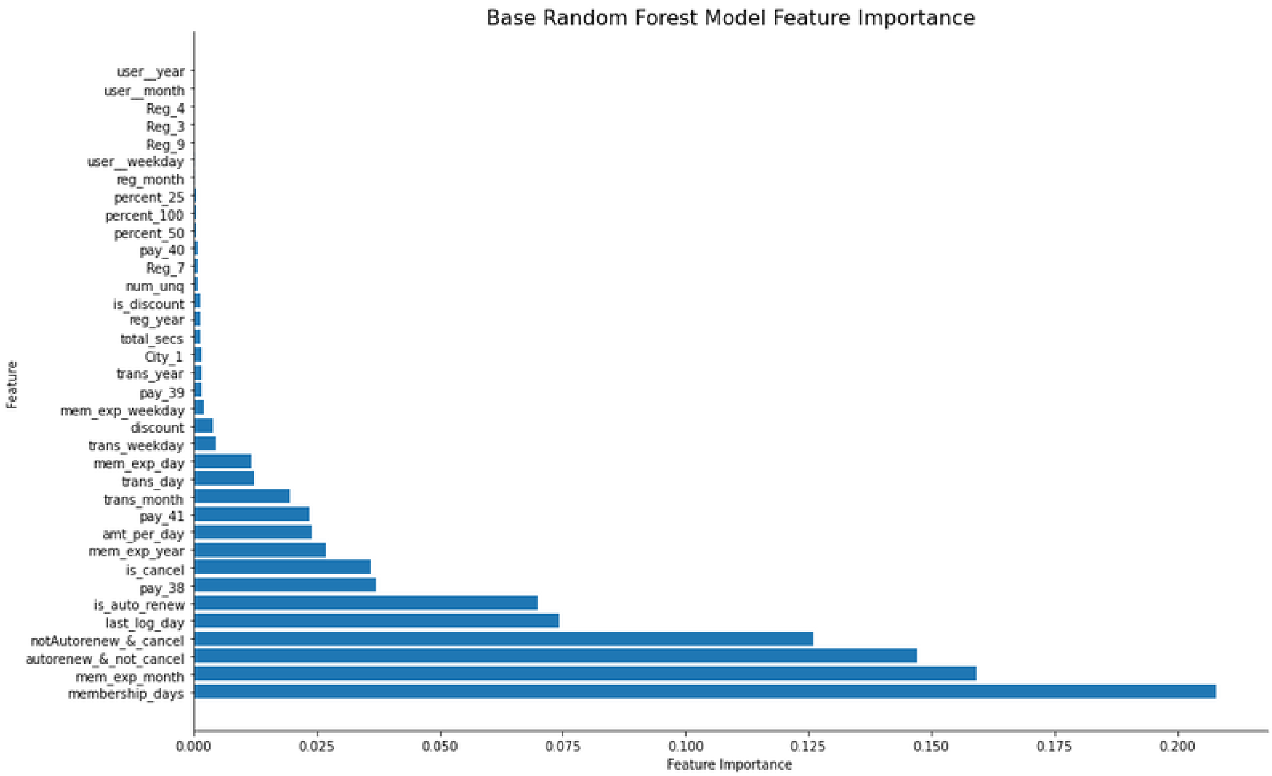
The optimal Random Forest Model has maximum performance for a max depth of 9 and 10 estimators.

Random Forest Model: ROC-AUC Curve

Validation AUC Score:0.99



Random Forest Model: Feature Importance



Concluding Remarks: Actionable Insights

The Random Forest Model feature importances are consistent with the findings of the Logistic Regression Model odds ratio. The membership duration has the biggest impact on churn prediction. The Random Forest Model shows that the combinations of autorenewing and not cancelling an existing membership as well as not autorenewing and cancelling an existing membership are top important predictors of churn. As expected, these specific combinations give more information and lead to a better forecasting of churn. It is also interesting to note that the original features autorenewal and cancellation do not drive the prediction as much as their combinations do and offer less insight.

An increase the membership duration, which is the highest driver of churn according to both models, can reduce the likelihood or odds that a user will churn. The records where the users last logged in to listen to music would show if there is any indication of decreased user app activity. This behaviour would signal a red flag that the user might cancel soon, so marketing should target advertising specials to these users in order to retain them. If the user cancels, marketing can offer new membership sign up promotions or plan discounts as discount also drives churn prediction.

Random Forest Model Final Test Performance Metrics

	Precision	Recall	F1-Score
No Churn	0.98	0.99	0.98
Churn	0.86	0.81	0.84

Test AUC Score: 98.7%

Churn Rate Prediction Impact on Revenue Loss

Churn Prediction = **$ 1.9 million** in lost revenue within the next 30 days

By predicting the churn rate of 1.3%, KKBox can expect a revenue loss of $1.9 million if customer retention is not improved. These findings can help KKbox gain a better understanding of their future expected revenue.

Additionally, KKBox can use these results to identify and improve upon areas where customer service is lacking. Their marketing team can understand what preventative steps are necessary to ensure lost revenue is minimized. Particularly, they can focus on the most important features from our Random Forest Modelling results to make improvements in customer lifetime value as well as retention and further reduce the churn rate and revenue loss.