KKBox Churn Prediction Project

By Marshall Lee
2/2/2023

Glossary

Problem Statement	Data Wrangling	Exploratory Data Analysis	Model Building and Evaluation	Conclusion
What are the features most associated with user churn, and how can they be improved to reduce churn rate?	Clean and normalize our dataset, which contains a user's listening habits, transaction history, and some personal information.	Identify relationships between features and target variable. Visualize trends using bar graphs, scatter plots, etc.	Create test/training sets and compare the performance of three models: Random Forest, Logistic Regression, and XGB. Tune hyperparameters and evaluate metrics of final model.	Model deployment and final recommendations to stakeholders.

Problem Statement

- KKBox is Asia's leading music streaming service, holding the world's most comprehensive Asia-Pop music library with over 45 million tracks.
- The purpose of this project is **determine which features are most associated with user churn,** and to build a model that can accurately predict churn rate
- A user is labelled as "churn" if he or she did not make a valid subscription within 30 days of the current membership expiration (March 2017 in this case)

Step 1: Data Wrangling

Data Wrangling

- The dataset contains over 700,000 entries and 20 different features, including our target churn variable which is listed as “True (1)” or “False (0)”
- Data includes information about a user’s transaction history, listening habits, and registration/expiration information
- Dataset contained null values, duplicate entries, typos, and in some cases, inaccurate data
    - User transactions that took place after April 2017 were purposely omitted by the host of the competition
- Contained no obvious outliers
- Data source and details can be found here

Data Wrangling cont.

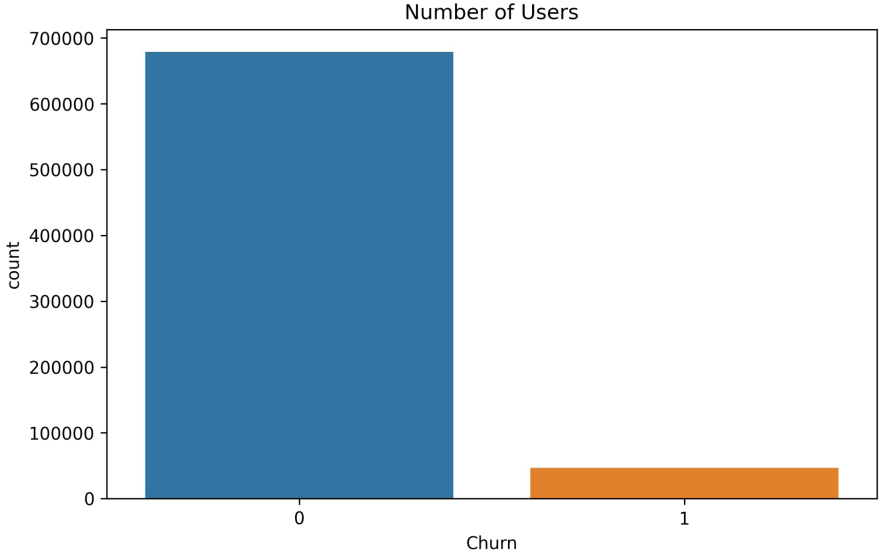
- Dealing with dirty data:
  - All entries with null values were removed
  - For both duplicate and non-duplicate entries, the total membership duration and latest transaction dates were calculated using the sum and max, respectively
    - Max values were then calculated for payment and registration information
    - Average time between transactions was also calculated

Step 2: Exploratory Data Analysis

Exploratory Data Analysis

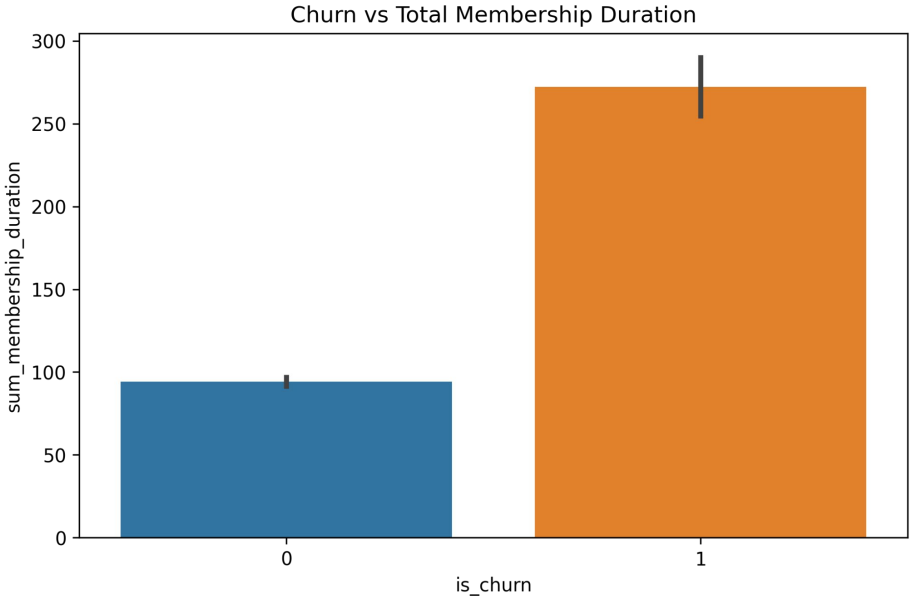
- In March 2017, 46603 users out of 725,723 users churned. This equals a **6.4% churn rate**

- In the next slides, we will explore each of the top features associated with user churn



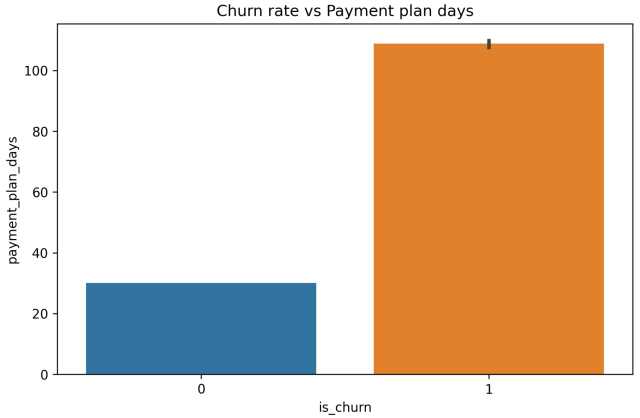
Tenure

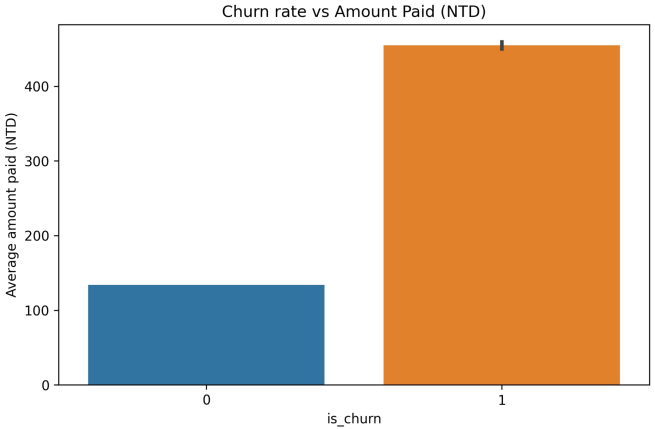
- At first glance, **total membership duration** seems to be highly correlated with user churn
- On average, churned users are subscribed for about 6 months longer than non-churned users.
- A majority of users opt for monthly memberships. However, churn rate increases by 4% for users having memberships longer than 32 days.



Payment details

- Further exploration reveals payment features to also be highly associated with churn rate - namely, **payment plan days** and **amount paid**
- On average, churned users paid 300 Taiwan dollars (~ $10 USD) more and opted for longer (~90 days) payment plans than non-churned users
- Payment plan days and amount paid features have correlation coefficients of 0.50 and 0.49 to churn rate, respectively
- Both features have a p-value of less than 0.001

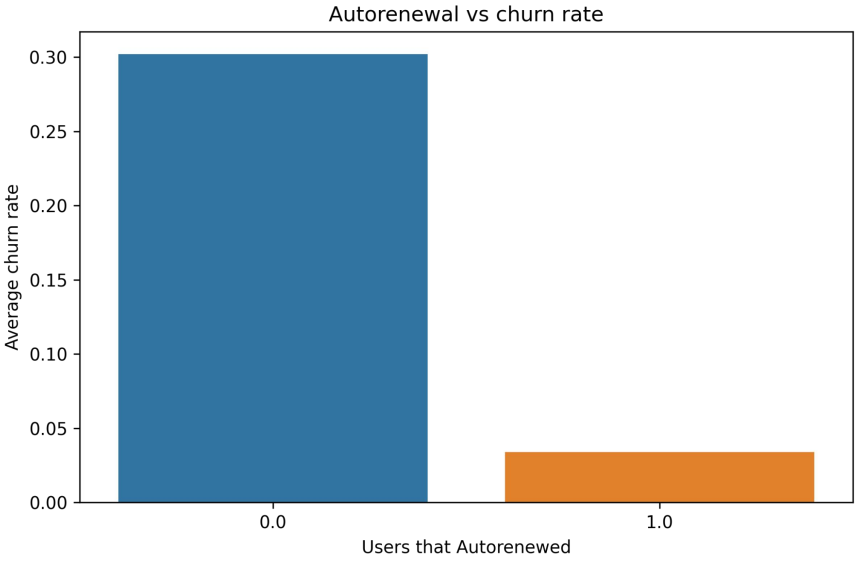




Auto-renewal

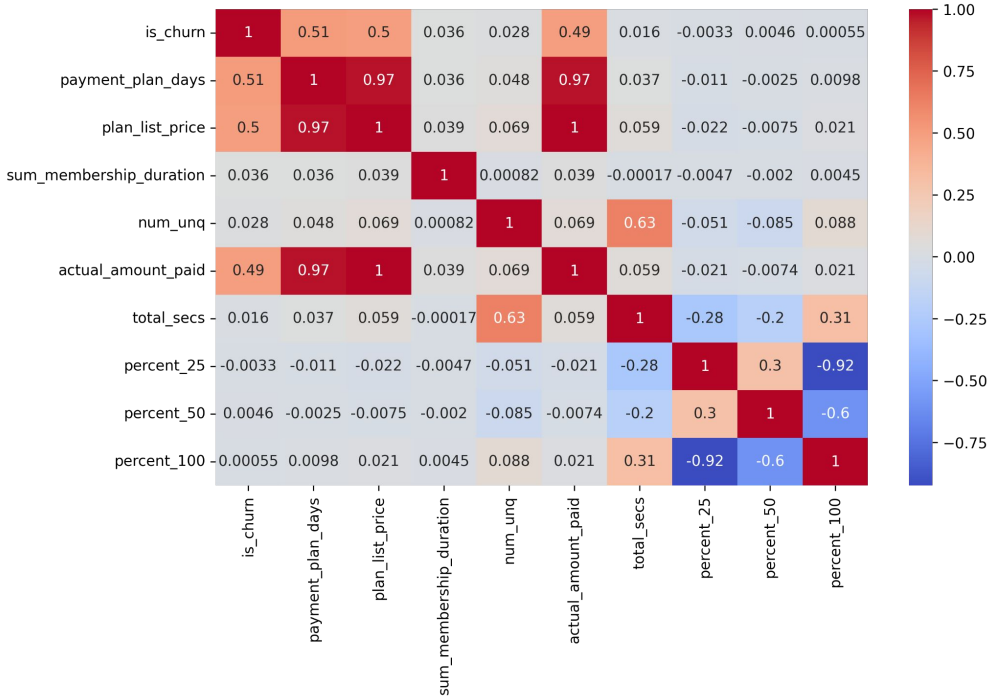
- Users that chose to **automatically renew** their subscription rather than to manually renew it were 25% less likely to churn

- Had the fourth highest correlation coefficient of 0.34



Correlation of numerical variables

- After analyzing our top features, we see that some features are correlated with each other: payment plan days, plan list price, and actual amount paid
- Plan list price and actual amount paid showed the highest correlation (0.99). Therefore plan list price was removed from our feature selection.



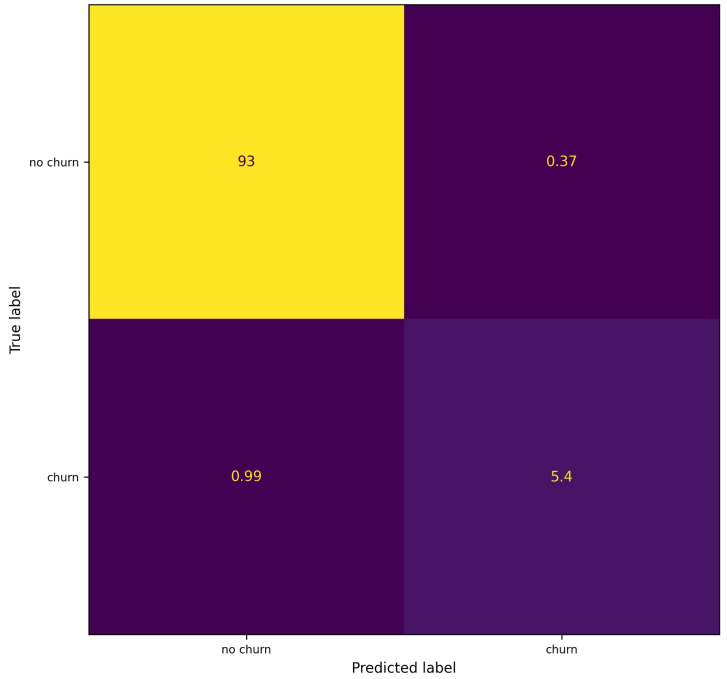
Step 3: Model Building and Evaluation

Model Building

- Machine Learning models built:
  - Random Forest
  - Logistic Regression
  - Extreme Gradient Boosting
- Besides the date-time data-types, the dataset was already label encoded. After the date-time objects were encoded, all features were scaled using StandardScaler and then split into training and test sets
- Top model will be chosen based primarily on its Precision, Recall, and F1 scores
  - Since the main focus is user retention, incorrectly labelling a non-churned user as churn will not be as costly as mislabelling a churned user and non-churn.
  - We want our model to identify as many churn users as possible

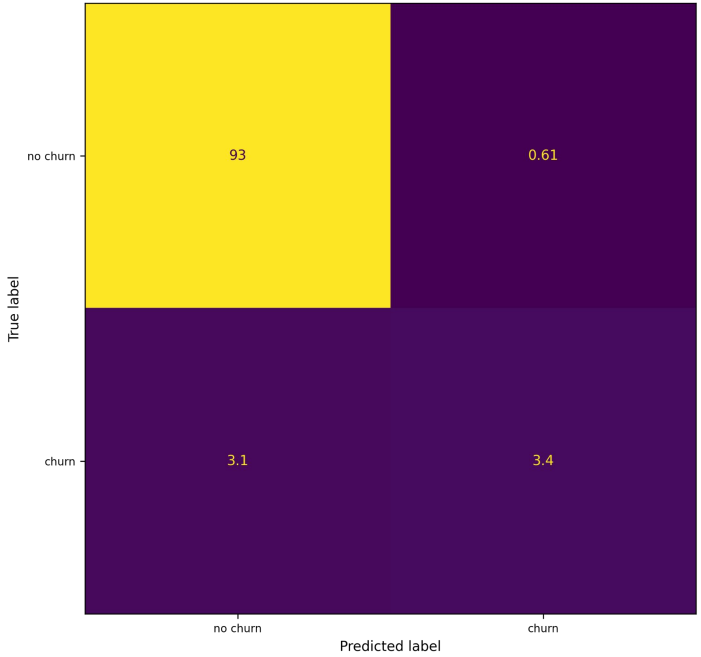
Random Forest Performance

- Random Forest model was tuned using RandomizedSearchCV. Optimal hyperparameters were:
  - N-estimators: 100
  - Max depth: 20
- Precision: 94%
- Recall: 84%
- F1: 89%



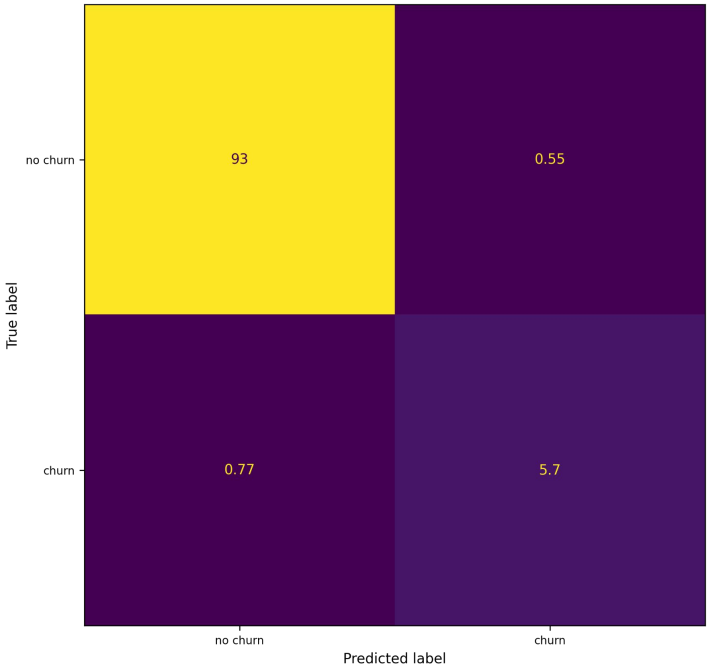
Logistic Regression Performance

- Logistic Regression model optimal hyperparameters (RandomizedSearchCV):
  - Solver: Newton-cg
  - C: 10
- Precision: 84%
- Recall: 53%
- F1: 65%



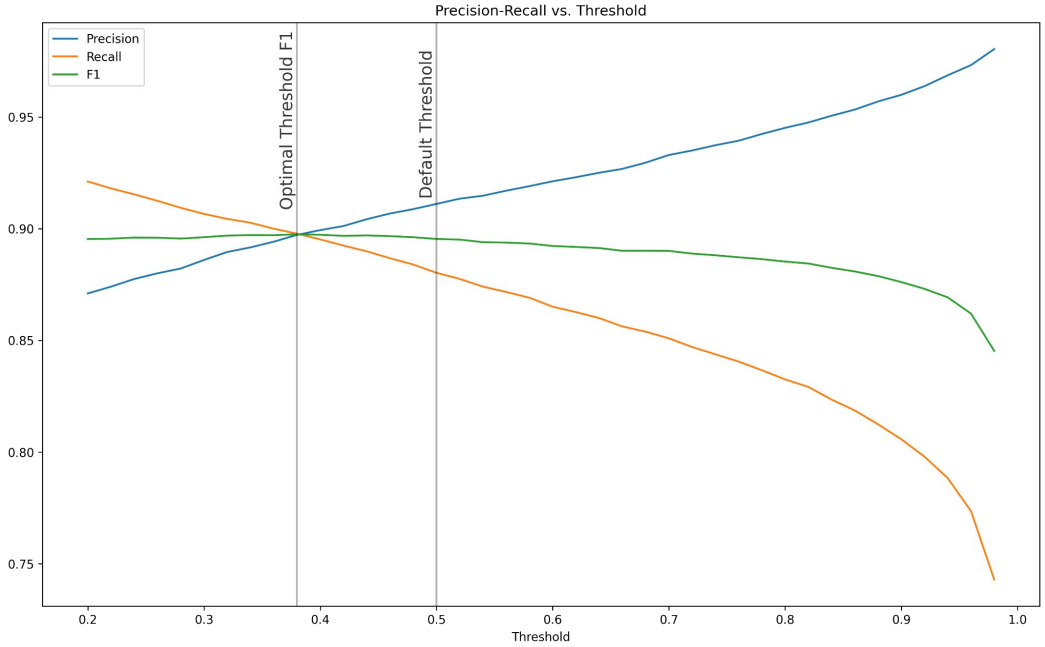
Extreme Gradient Boosting Performance

- XGB model optimal hyperparameters (RandomizedSearchCV):
  - N-estimators: 100
  - Minimum child weight: 1
  - Max depth: 20
- Precision: 91%
- Recall: 88%
- F1: 90%



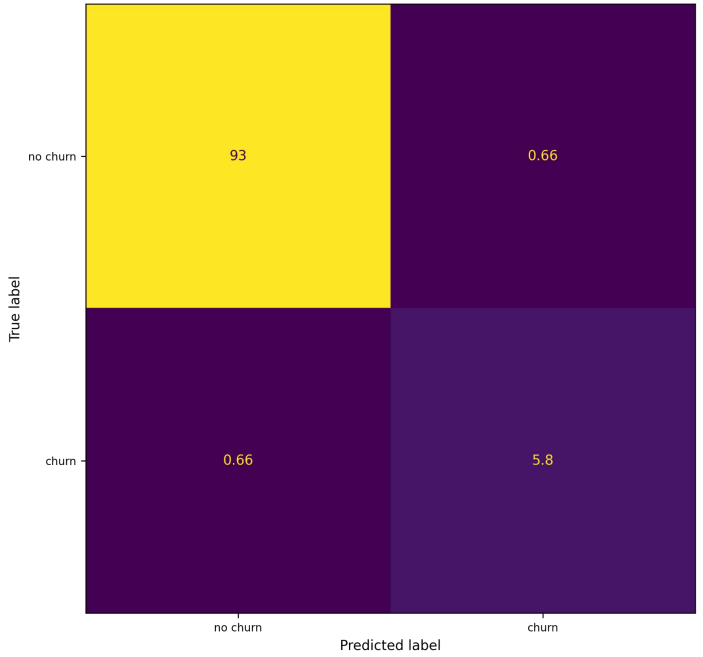
Threshold adjustment

- Due to imbalance in our dataset, we also want to adjust the threshold. A **0.38** threshold maximizes PR/F1 scores and is therefore the optimal threshold



Final Selection and Evaluation

- Due to its slightly higher recall and F1-scores, the Extreme Gradient Boosting model wins by a small margin
- To further adjust for the class imbalance, threshold was changed from the default to 0.34
- Summary of our final model metrics:
  - Precision: 91%
  - Recall: 90%
  - F1: 90%
  - Optimal threshold: 0.34



Conclusions

- It is clear long-term users seem to churn more often. Auto-renewal is also strongly correlated with churn-rate. Therefore, focusing on retaining long-term users will be the very effective at reducing churn rate
- Extreme Gradient Boosted model helps predict whether a user will churn or not based primarily off of how much the user is paying, their payment plan period, and their tenure
- Suggestions:
    - Gather more data in the form of customer surveys. Brainstorm solutions based on this customer feedback. Solutions include a new pricing model, discount offers, and/or loyalty programs
    - Develop strategies that encourage users to automatically renew their subscription in the form of notifications, PSAs, etc.
- If time permits, we can choose to gather churn data for the following months to identify monthly/yearly trends we may be missing with our current dataset

Thank you!

Thanks to Ben Bell from Springboard who was my mentor for this project
Full project notebook and report can be found here