Churn Prediction for KKBox

Group 2: Xinyao Guo, Qingyun Lu, Peilin Qiu, Sijian Xuan, Yi Zhang

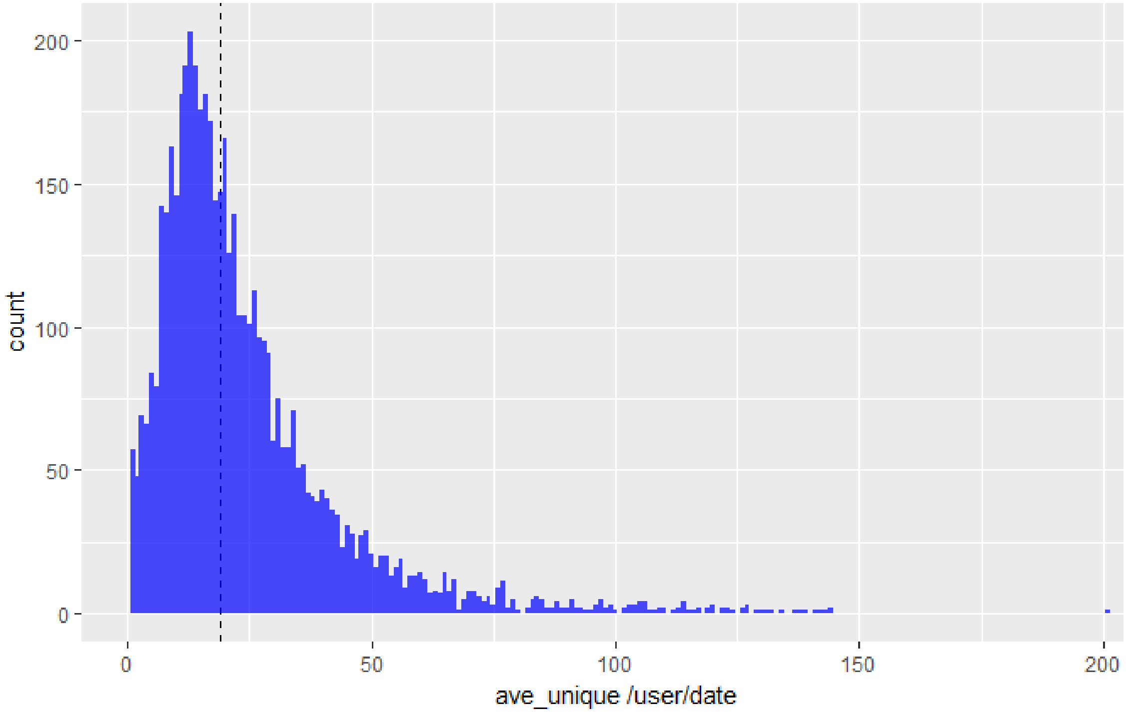
Overview

- Introduction
- EDA
- Model Fitting & Selection
- Future Improvements

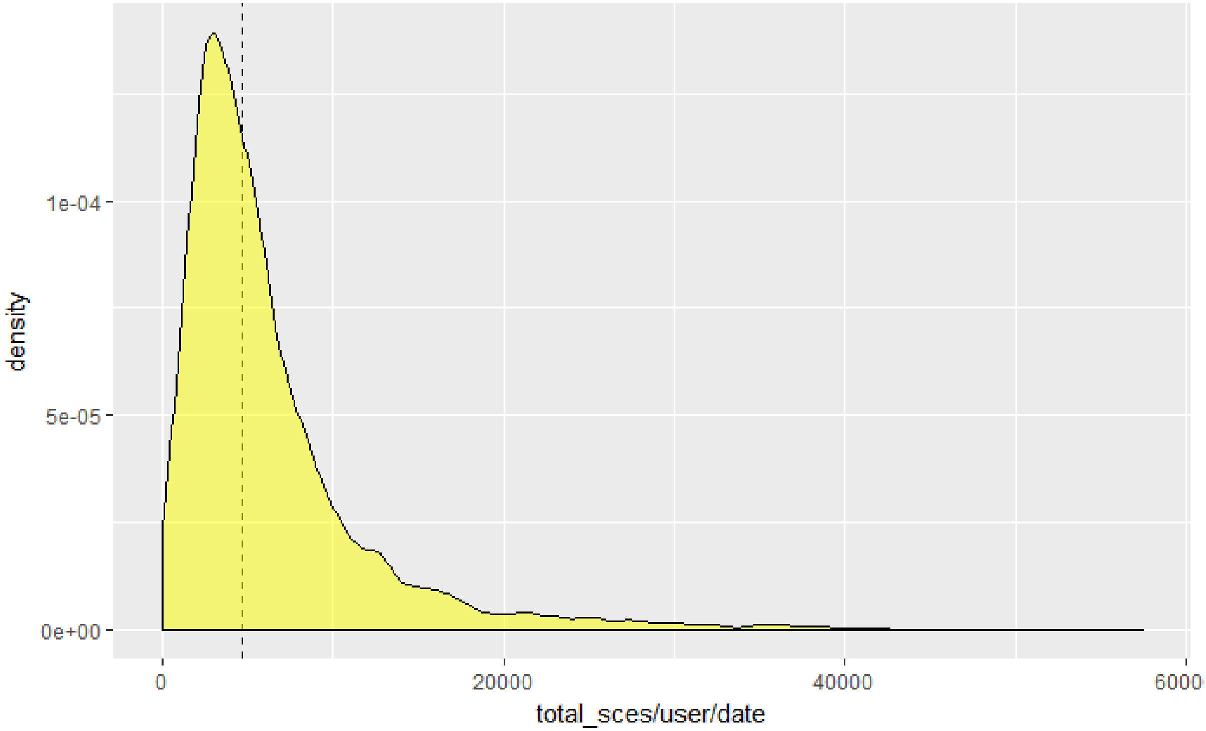
Introduction

- **Goal:**
  - Understand important drivers of the retention rate for popular musical Apps
  - Build predictive model to forecast whether a user will churn after subscription expires
- **Data:**
  - User information at member, transaction and user logs level.
  - Merge data by id and select 5000 common users
  - Features Engineering: Select and create certain features based on EDA
  - Feature Importance: Random Forest, XGBoost
- **Models:**
  - Baseline Model: SVM
  - Advanced Models:
    - Random Forest, GBM, XGBoost, ADABoost, Lasso

Exploratory Data Analysis
--- User logs

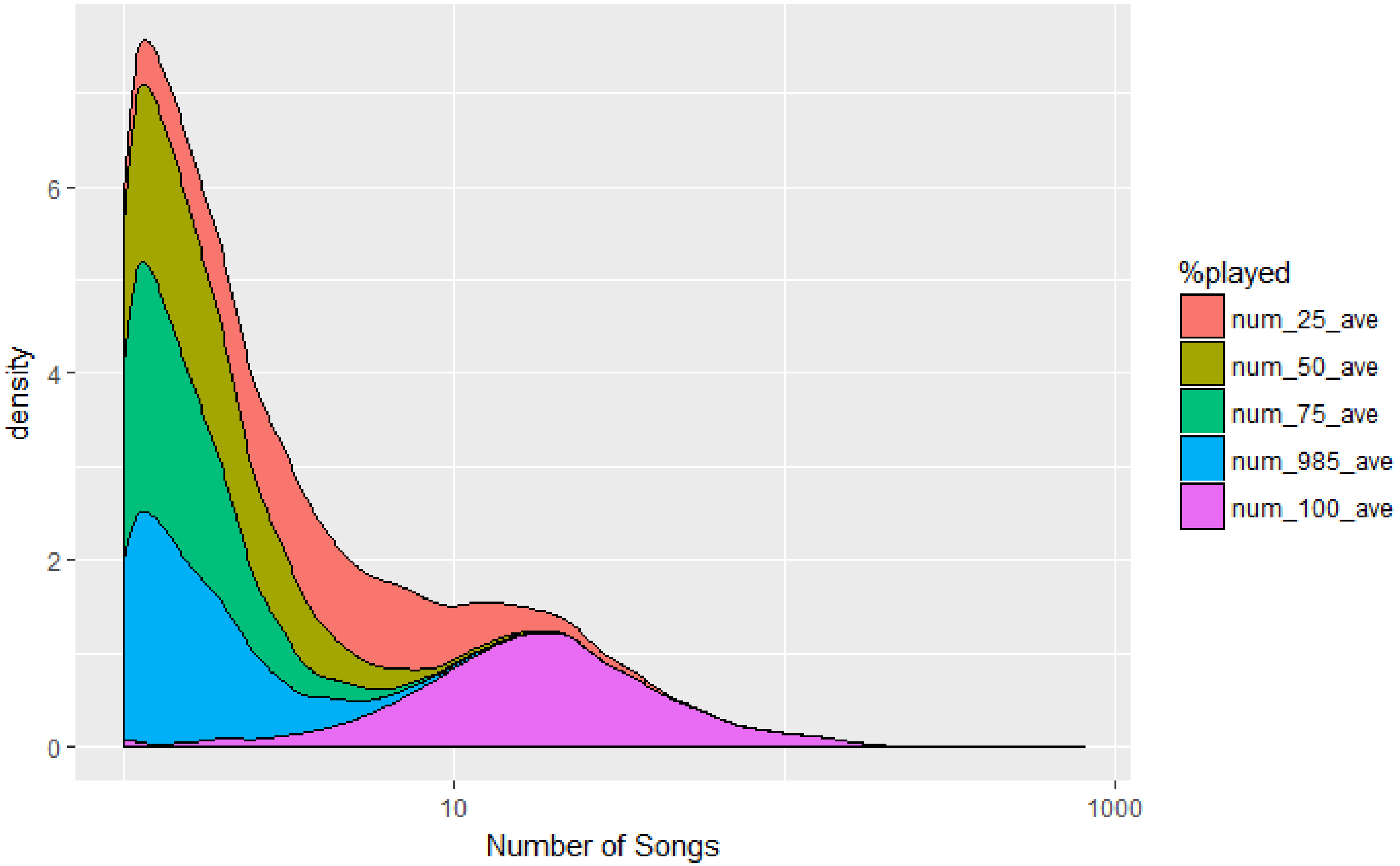


Number of unique songs per day



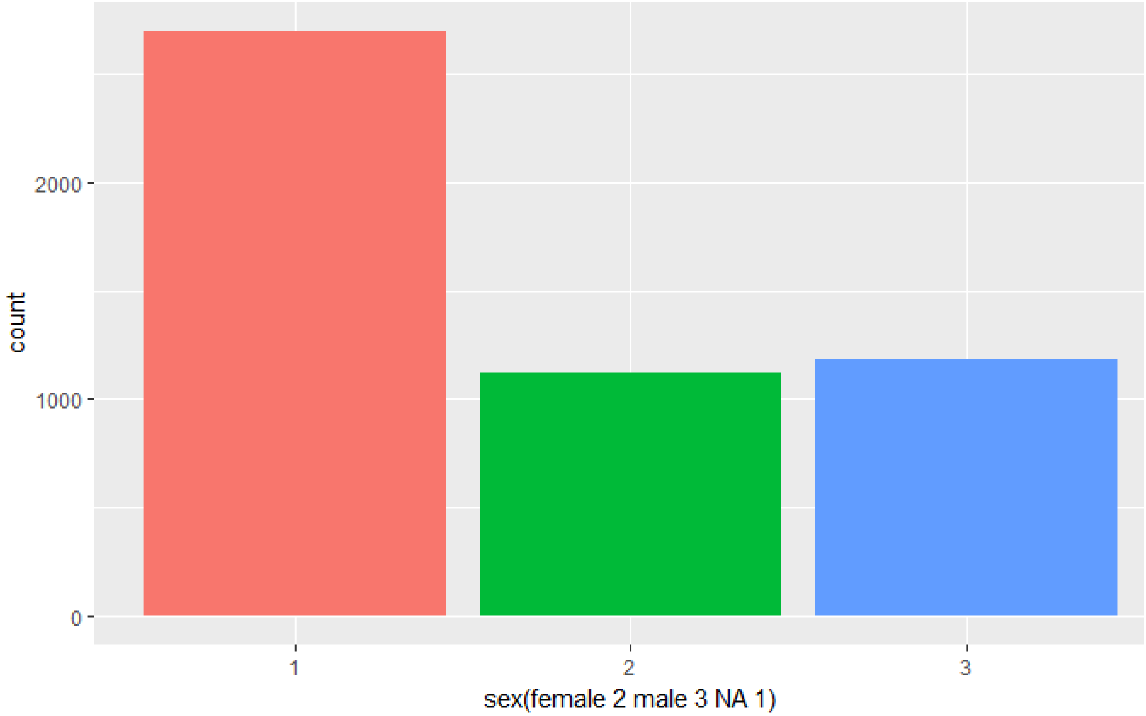
Total listening time per day

User logs

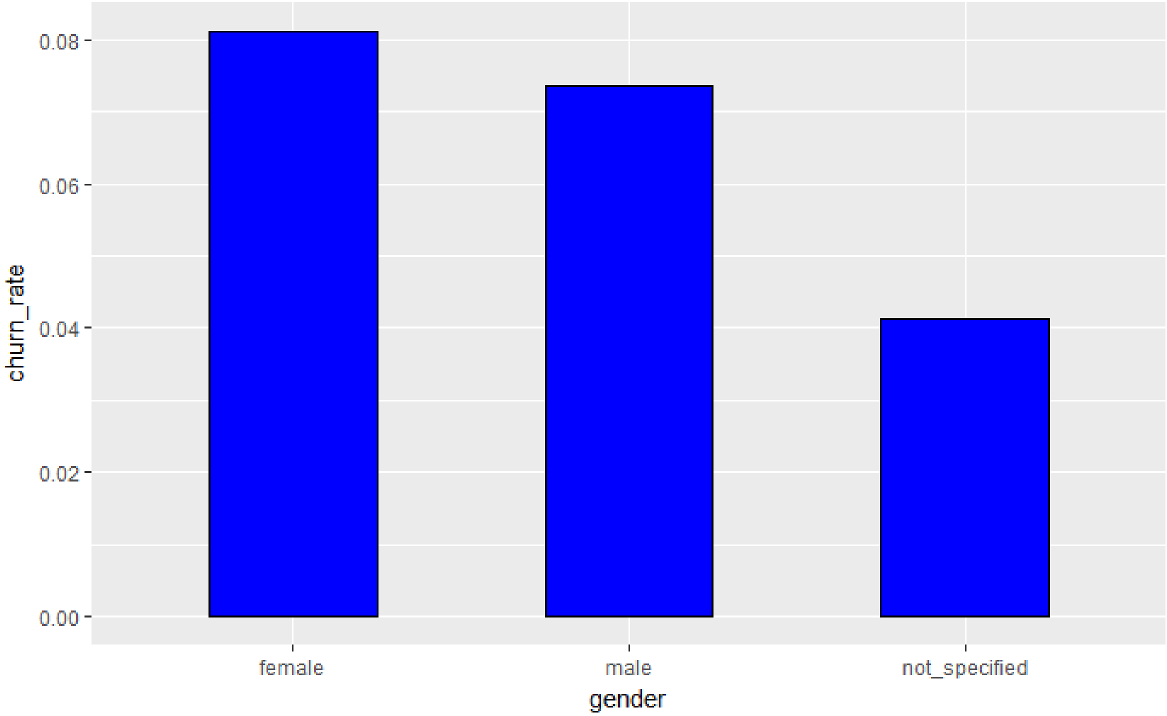


Distribution of number of songs played

Member

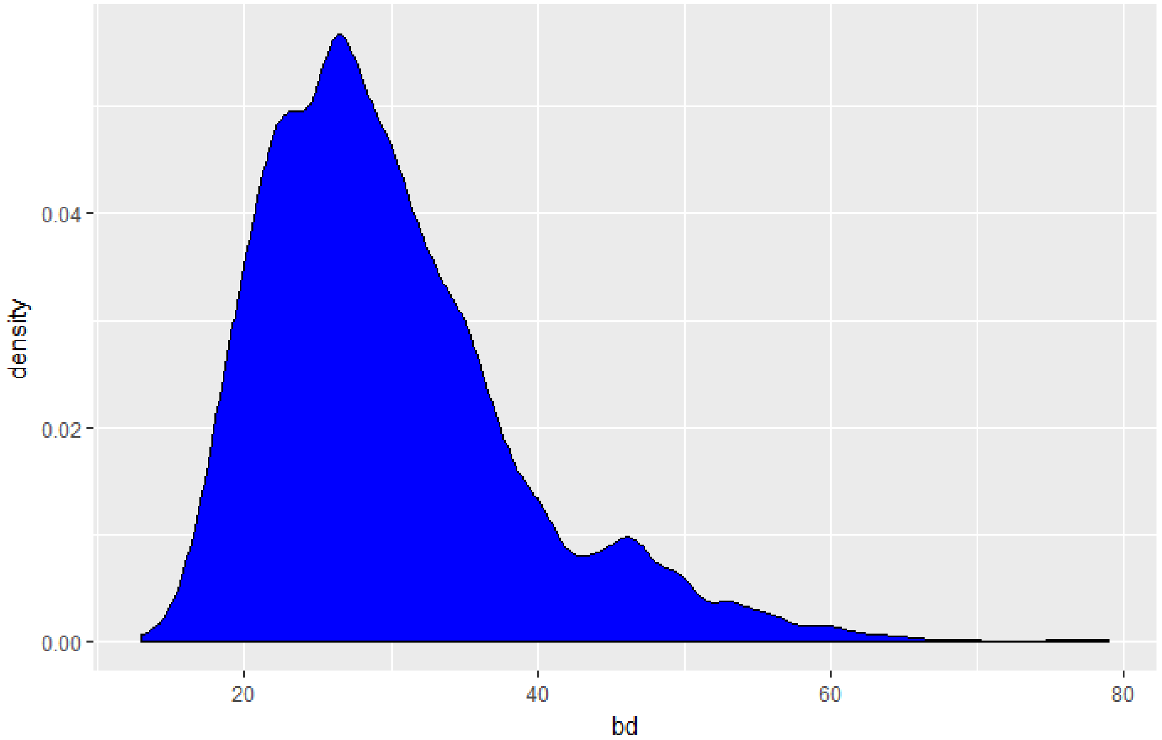


Gender

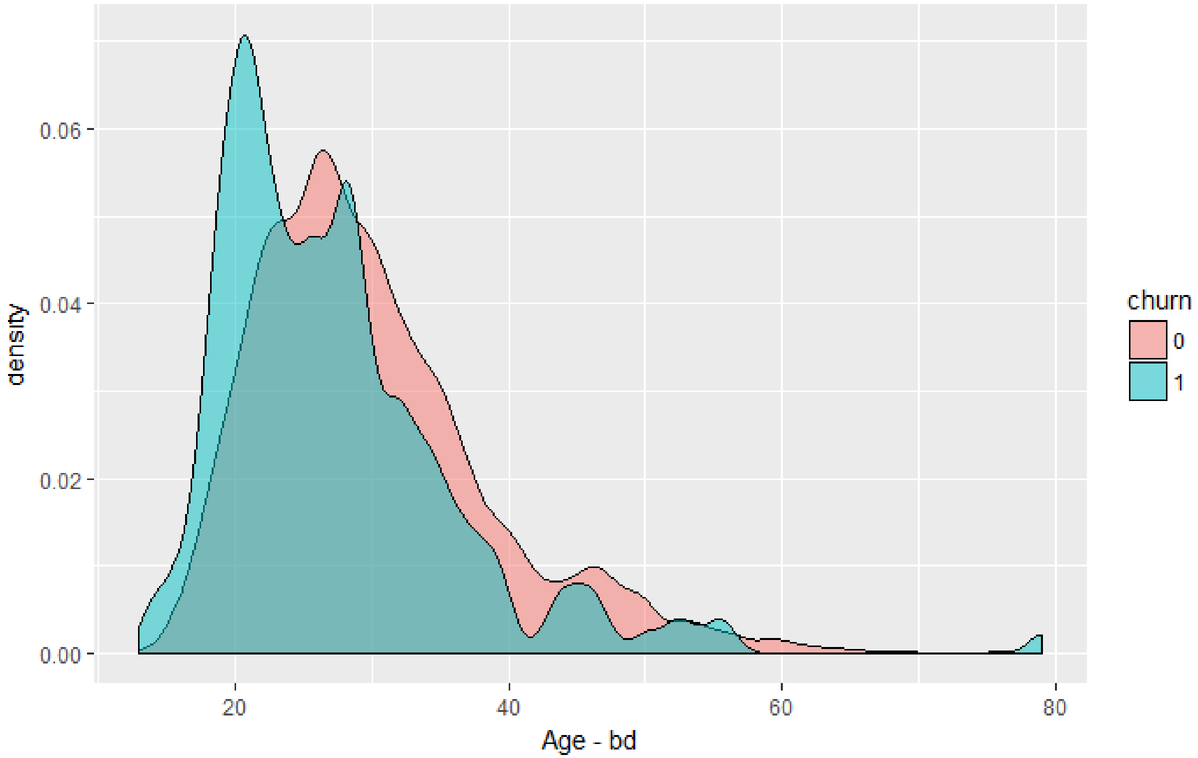


Relationship between gender and churn rate

Member

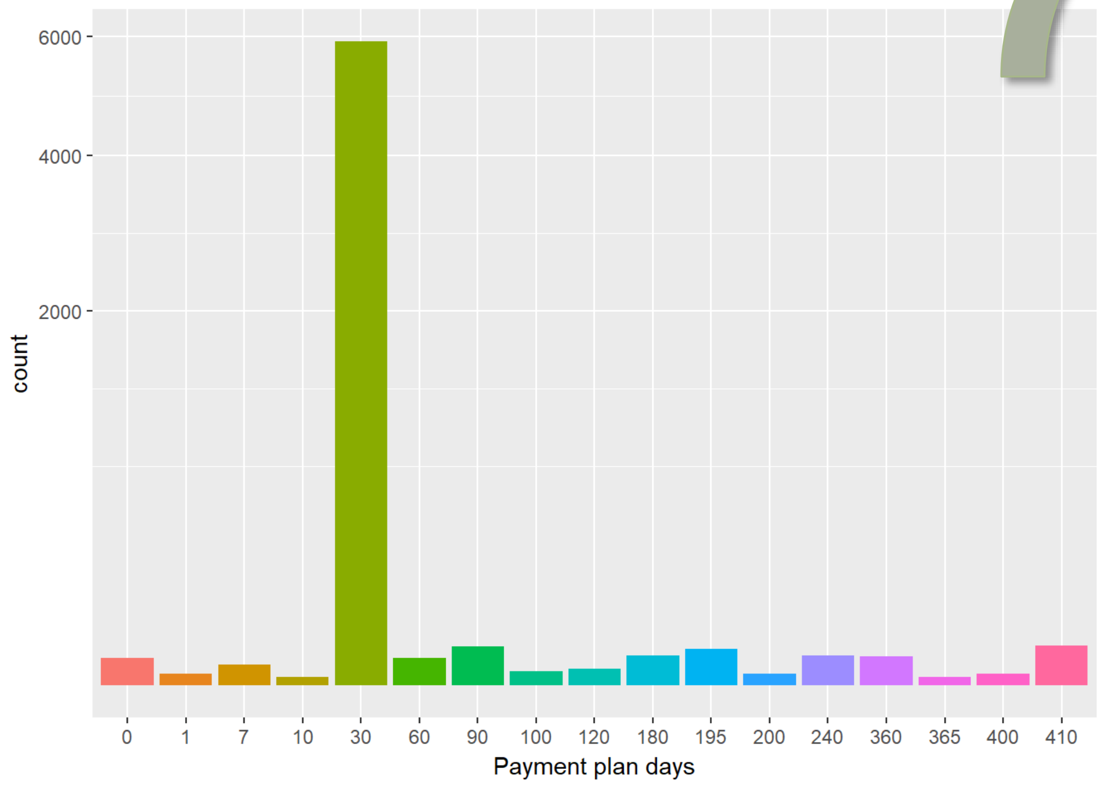


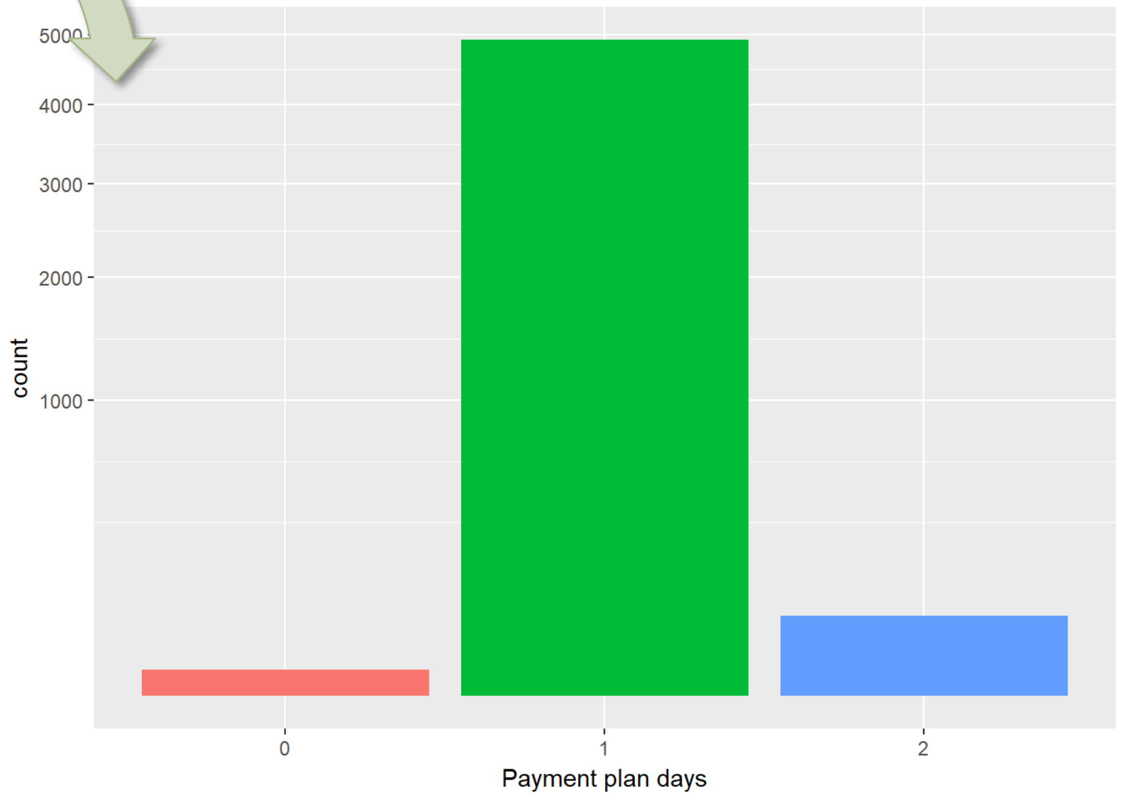
Distribution of age



Age & is_churn

Transaction

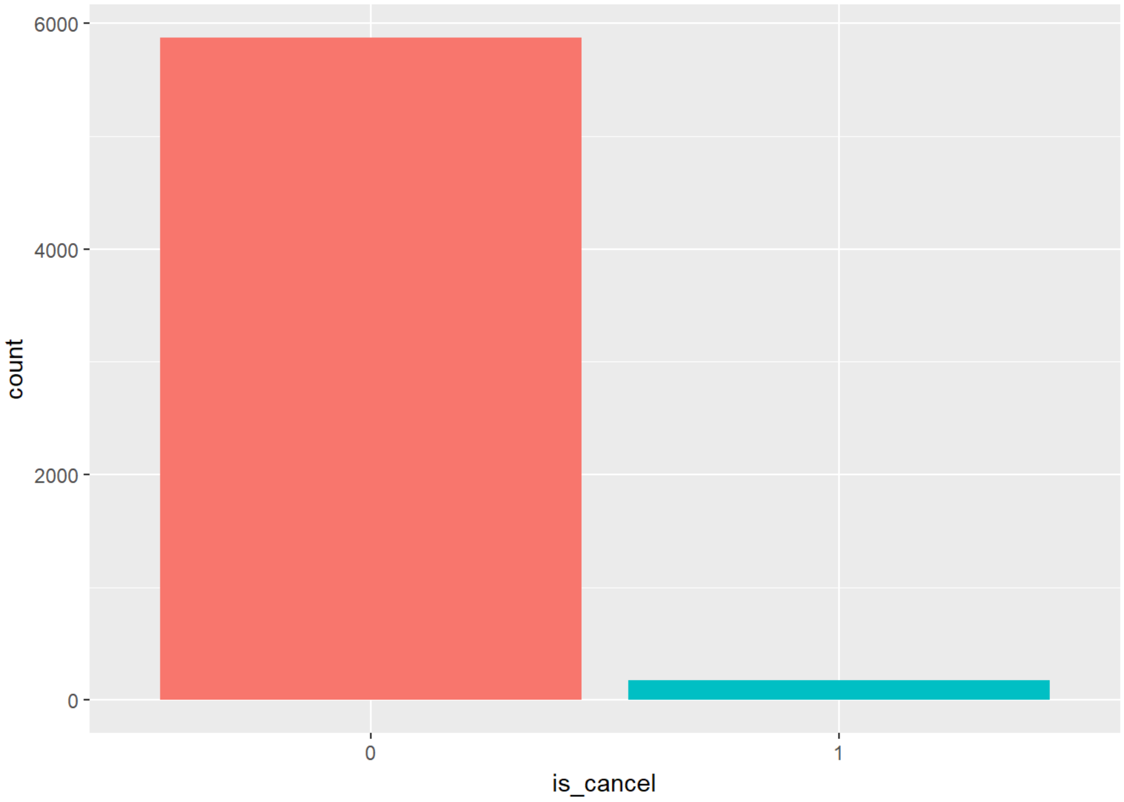




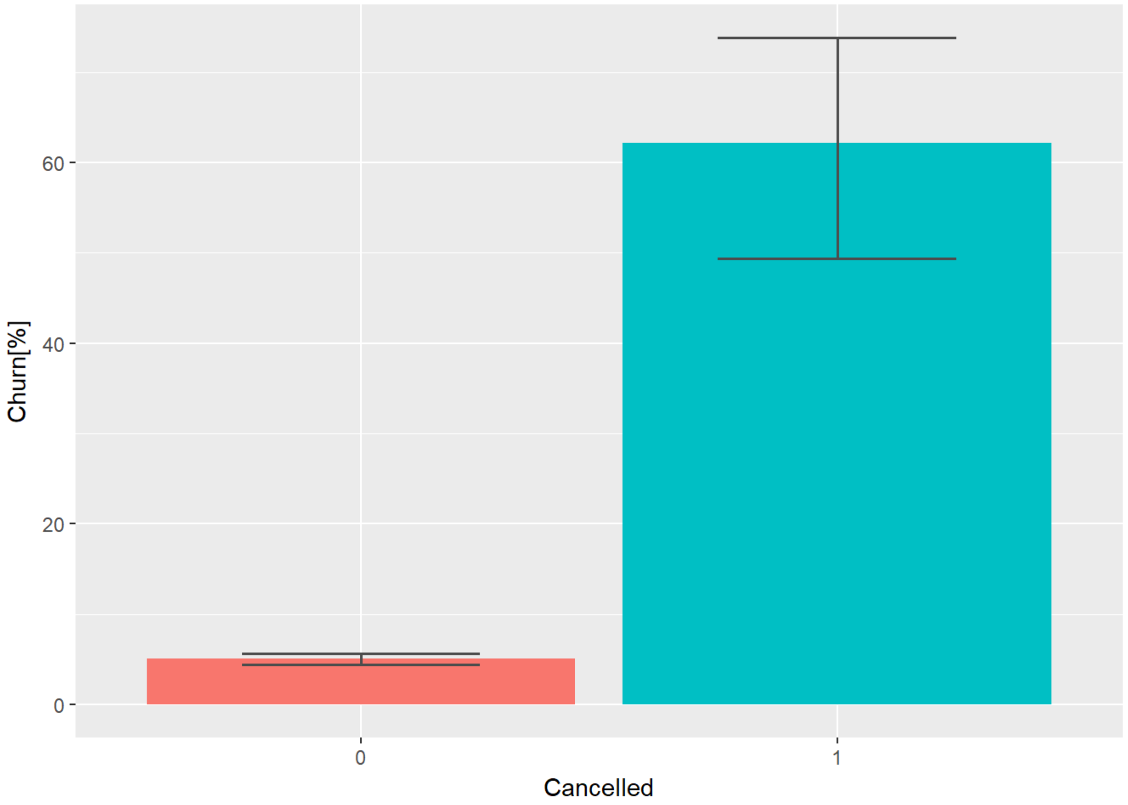
Payment plan days

0: <30, 1: 30-90, 2: >90

Transaction

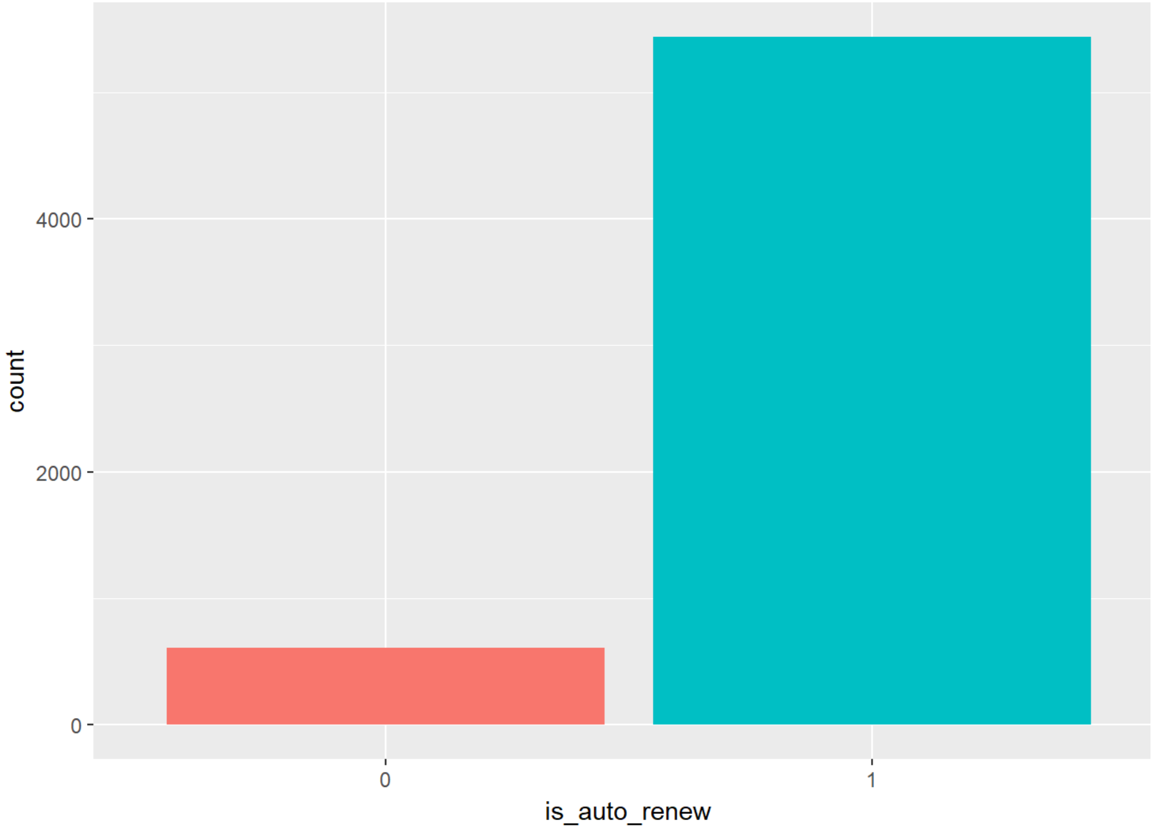


Is_cancel

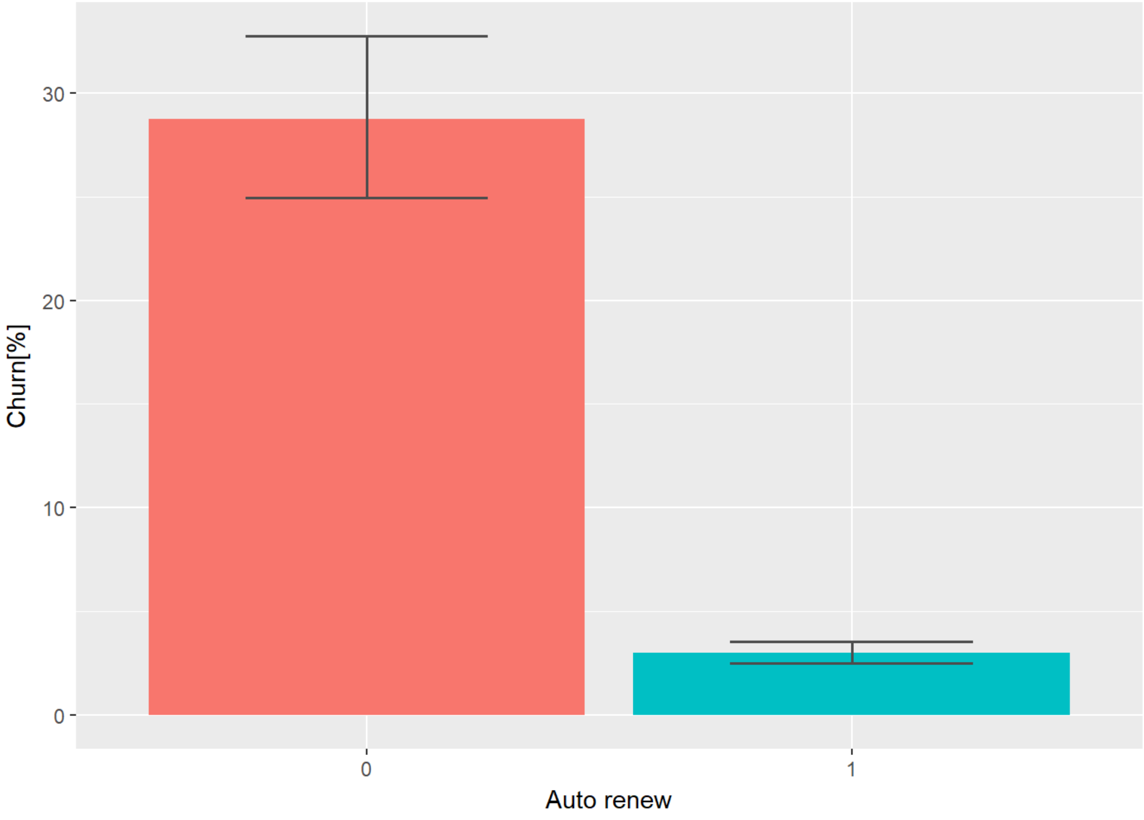


Is_cancel & churn rate

Transaction

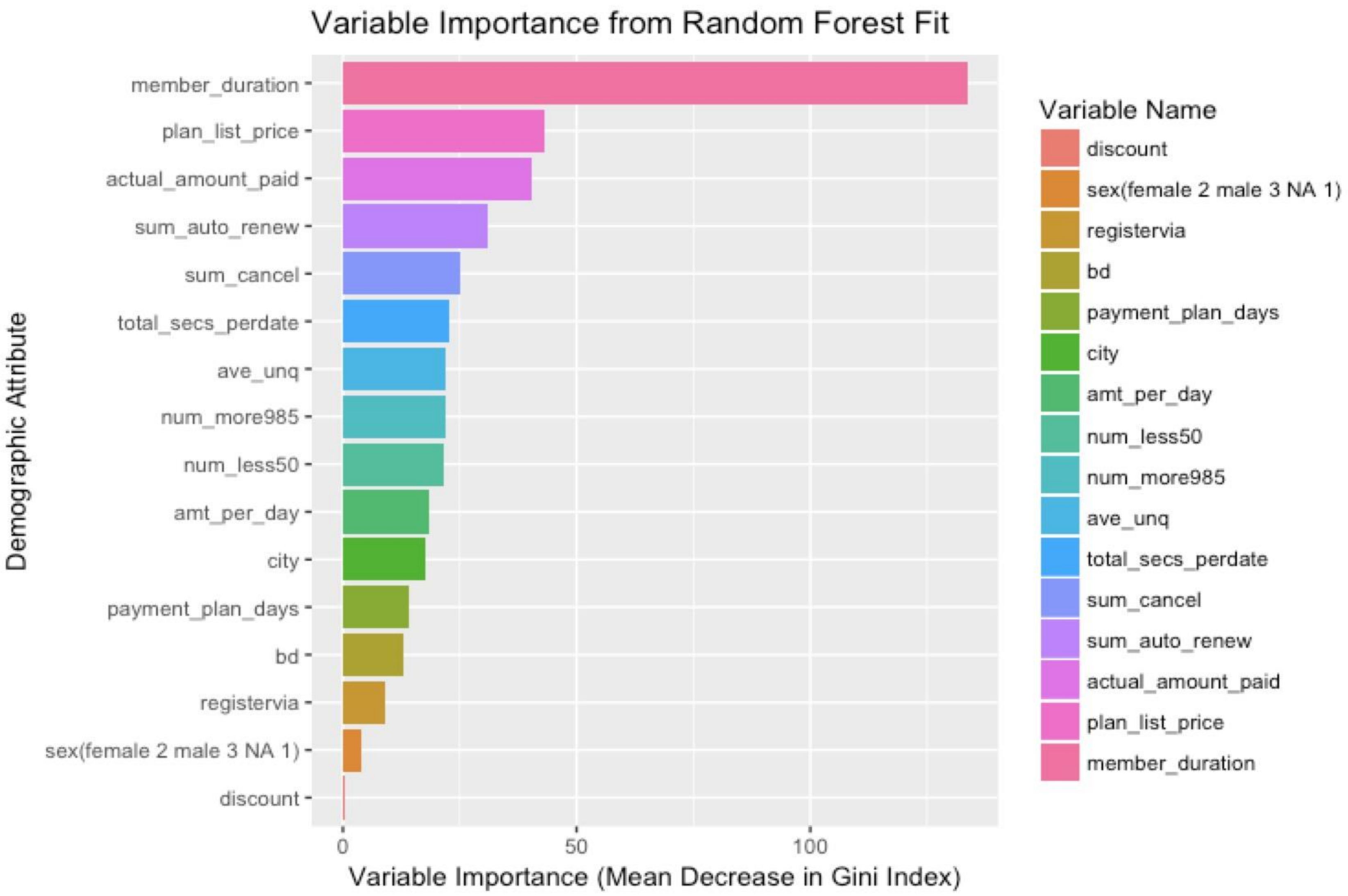


Is_auto_renew



Is_auto_renew & churn rate

Feature Importance – RF



- A low Gini (i.e. higher decrease in Gini) means that a particular predictor variable plays a greater role in partitioning the data into the defined classes.
- Gini impurity index

$$G = \sum_{i=1}^{n_c} p_i(1 - p_i)$$

Where n_c is the number of classes in the target variable and p_i is the ratio of this class.

Evaluation

To deal with the imbalanced labels (90% of the users will not churn), we use Log Loss to evaluate our prediction:

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1-y_i)\log(1-p_i))$$

where N is the number of observations, yi is the binary target, and pi is the predicted probability that yi equals to 1.

Model Fitting & Selection

- 1. Train-Test randomly split:
  - 80% training set VS. 20% testing set
  - Metric: Log Loss
- 2. Tune Model—Grid Search with CV
  - Baseline model: SVM
  - Advanced Model:
    - XGBoost
    - AdaBoost
    - Lasso
    - **Random Forest**
    - GBM

Models		Test Error	CV Error
Baseline	**SVM**	0.1572	0.1678
Advanced	**XGBoost**	0.0749	0.1145
	AdaBoost	0.1166	0.1363
	Lasso	0.1446	0.1689
	Random Forest	**0.0620**	**0.0822**
	GBM	0.0655	0.0866

- Conclusion: Random Forest is the optimal model in our case.

Deal with Potential Overfitting

- 1. Use Cross-Validation to select model
- 2. Further simplify our models:
  - Ignore certain dataset
    - 1. Ignore UserLog dataset—perform worse
    - 2. Ignore Member dataset—perform slightly worse → maybe ignore certain features is feasible
  - Ignore certain features
    - 1. Ignore Gender, Registration Method in the Member dataset; Ignore Discount in the Transaction dataset
    - Conclusion:
      - Better result!
      - Random Forest is still our best model

Simpler Models		Test Error	CV Error
Baseline	**SVM**	0.1318	0.1381
Advanced	**XGBoost**	0.0848	0.1084
	AdaBoost	0.1137	0.1438
	Lasso	0.1351	0.1642
	Random Forest	**0.0590**	**0.0906**
	GBM	0.0674	0.0876

- 2. Ignore Gender, Registration Method in the Member dataset; Ignore Discount, Actual Amount Paid in the Transaction dataset
- Conclusion:
  - Worse result
  - Still stick with the previous model

Simpler Models 2		Test Error	CV Error
Baseline	**SVM**	0.1304	0.1495
Advanced	**XGBoost**	0.0745	0.1103
	AdaBoost	0.1175	0.1321
	Lasso	0.1364	0.1609
	Random Forest	**0.0623**	**0.0836**
	GBM	0.0661	0.0907

Results & Future Improvement

- Final Model: Random Forest model on data with reduced features
- Future Work:
  - Separate models for the users who provide valid information and no information (ex. Ignore the member dataset for those who did not have complete information and use a different model for them)
    - RF model for the group ignoring member dataset: Log Loss= 0.052
    - RF model for the remaining group with member dataset: Log Loss= 0.059
    - Need validation on a larger user base

Thank You!

Q & A