KKBox's Churn Prediction Report

Group 2 (Names):

Dhruva Kumar Kadiyala (dxk190028), Sanketh Reddy (spr150430), Madhuri Thorat (mxt200018), I-Ching Wang (ixw200000)

I. Introduction

KKBox is one of the most popular music streaming services in Asia with several subscription options to attract customers. Nowadays, companies are increasingly aware of the importance of subscription services, and the churn rate is a critical indicator to track the health of a subscription-based company. To be more precise, the company can take measures in advance by predicting the customer churn rate to retain customers consistently. Therefore, our goal is to help KKBox predict whether a subscriber will churn after his/her subscription expires.

II. Data Description

We obtained the data set from the Kaggle website (WSDM - KKBox's Churn Prediction Challenge). The datasets are composed of a user information dataset, transactions dataset, daily listening behaviors of a user dataset, and a training dataset. The datasets contain information from 6,769,473 users and include details about age, city, gender, churn data, payment method, length of membership plan in days, the number of songs played, etc. Our target variable is churn, and churn is defined as whether the user did not continue the subscription within 30 days of when his/her subscription expired.

Since the daily listening behaviors dataset reports daily metrics for each unique user, the original dataset contained duplicate values for the customer id column. To fix this duplicate customer id dilemma, we decided to group each unique customer id and get the sum of all of the column values for each unique customer id. This new and aggregate dataset contains all of the same information as before and gets rid of the duplicate customer id dilemma because each unique customer id is present only once.

To create the one dataset we used for our EDA, we joined the training dataset that kaggle provided to us and joined it with the users' information dataset, transactions dataset, and the new and aggregate users' listening behaviors dataset. Each dataset had the 'msno column' (customer id) and we joined each of the datasets by this column. The resulting dataset comprises of 22 variables and 725,722 observations.

The new dataset contains the following variables:

- msno: user id
- num_25: the total amount of instances each unique user listened to less than 25% percent of songs
- num_50: the total amount of instances each unique user listened to between 25%-50% of songs
- num_75: the total amount of instances each unique user listened to between 50%-75% of songs
- num_985: the total amount of instances each unique user listened to between 75%-98.5% of songs
- num_100: the total amount of instances each unique user listened to between 98.5%-100% of songs
- num_unq: the total amount of unique songs each unique user listened to songs
- total_secs: the total amount of seconds each unique user spent on listening to songs
- payment_method_id: payment method (33 Levels: 3 6 8 10 11 12 13 14 15 16 17 18 19 20 21 22 23 26 27 … 41)**

- payment_plan_days: length of membership plan in days
- plan_list_price: payment of membership plan in New Taiwan Dollar (NTD)
- actual_amount_paid: actual payment of membership in New Taiwan Dollar (NTD)
- is_auto_renew: whether the membership plan is auto renew or not (1, 0)
- transaction_date: transaction date (%Y%m%d)
- membership_expire_date: membership expiry date (%Y%m%d)
- is_cancel: whether the customer canceled the membership in this transaction or not (1, 0)
- city: city that customer lives (Levels: 1 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22)***
- bd: customer age (years)
- gender: customer gender (male, female)
- registered_via: registration method (Levels: 3 4 7 9 13)****
- registration_init_time: initial registration date (%Y%m%d)
- is_churn: whether the customer churned or not (1, 0)

**: Kaggle did not provide what each Payment Method id number actually refers to

***: Kaggle did not provide what each City number actually refers to

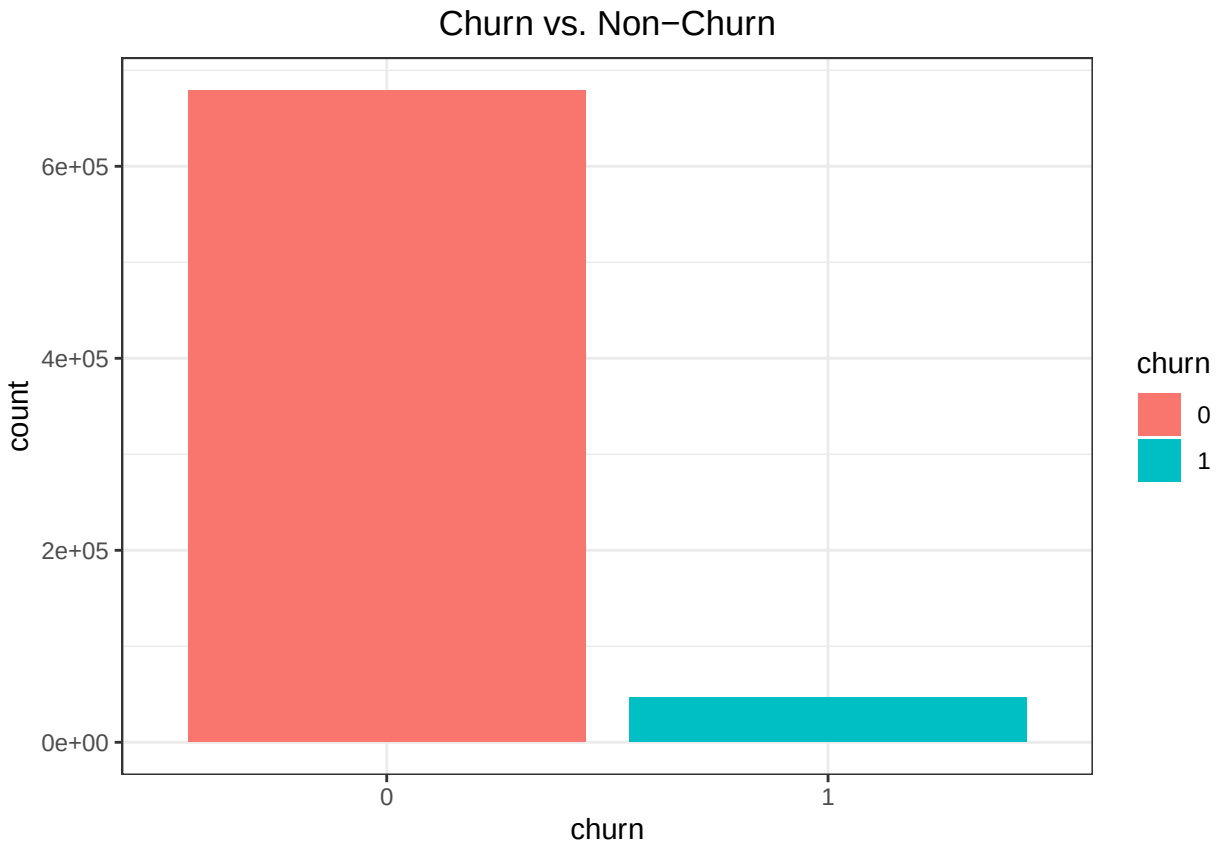
****: Kaggle did not provide what each Registration Method number actually refers to

III. Goal

The main issue that we are trying to resolve is figuring out the main and significant factors that are leading to customers churning. If we are able to figure out what these factors are, we can help KKBox reach out to certain customers to try and prevent them from churning after their subscription expires.

IV. Exploratory Data Analysis (EDA)

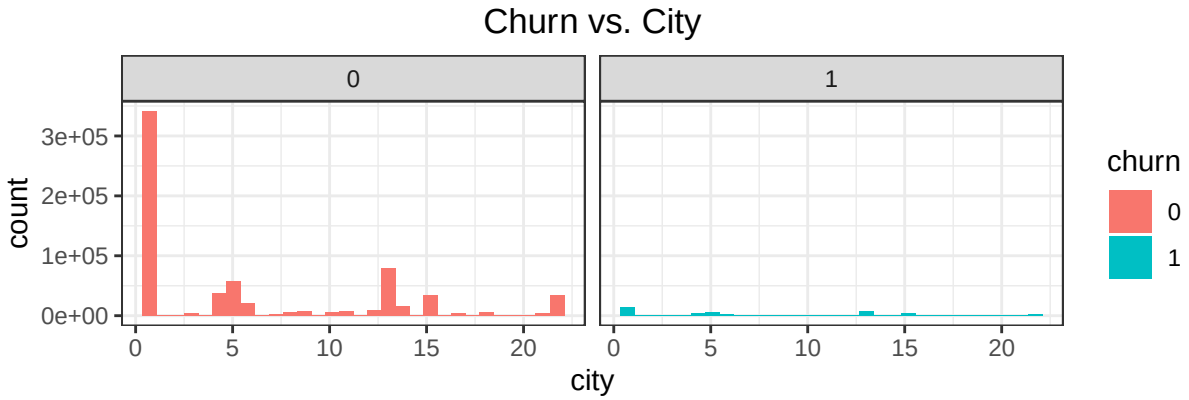
This is an imbalanced dataset. The plot and table below show that the customer churn rate in KKBox is low. The churn rate is approximately only 6% whereas the non-churn rate is approximately 94%.

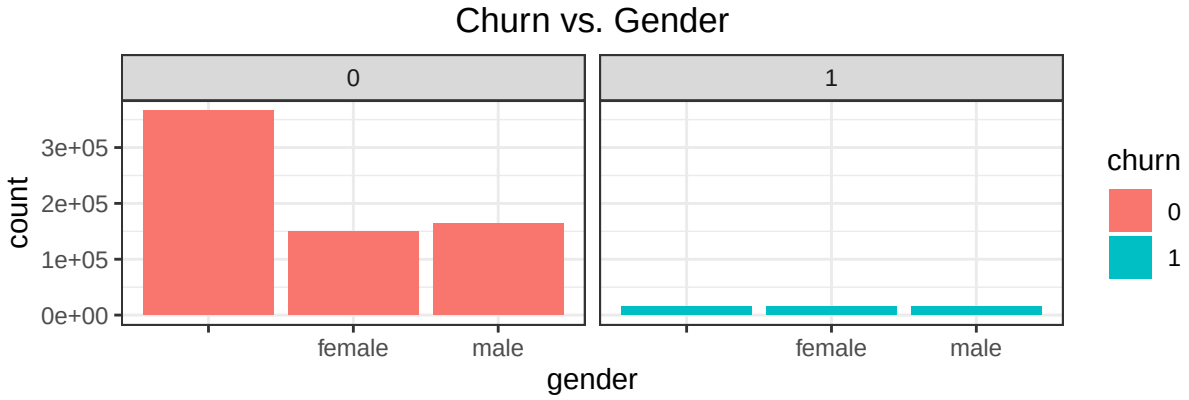



```
##
##     0     1
## 0.936 0.064
```

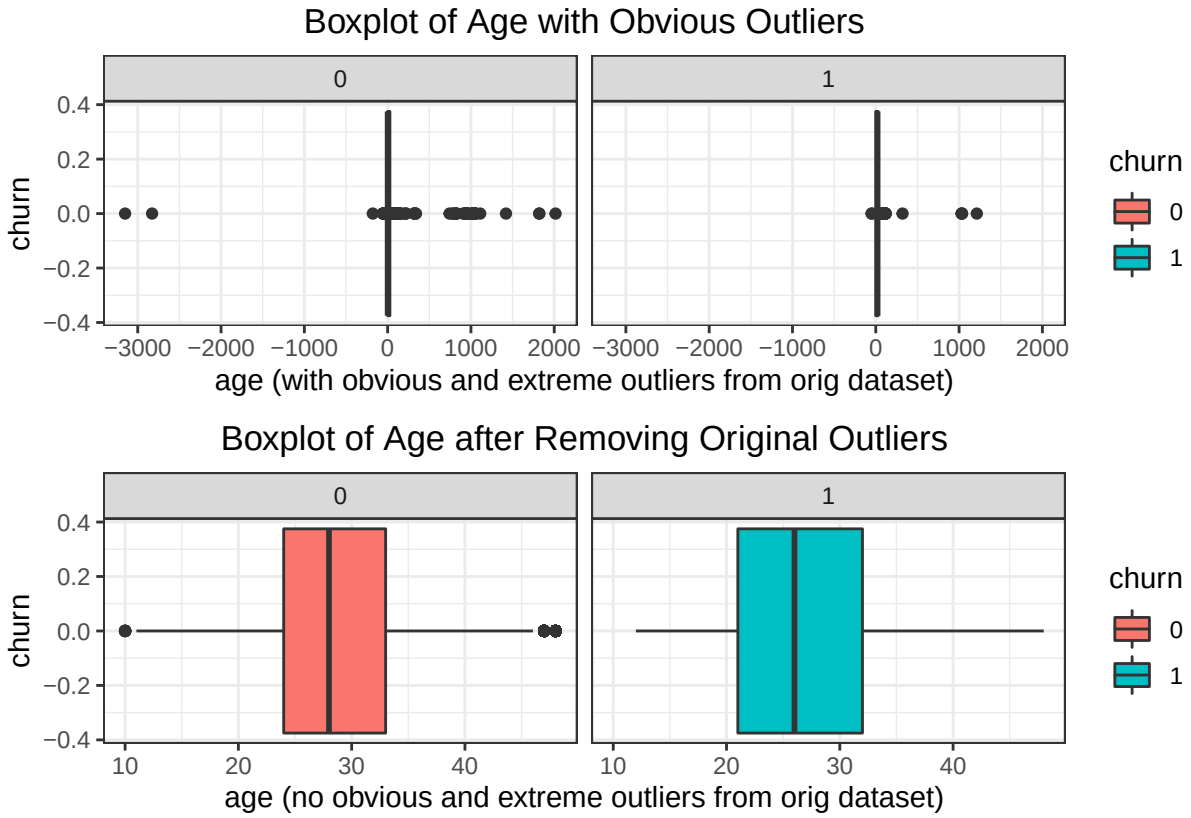
From the plots below, we can find two things. First, city 1* has the most number of users who did not churn. Second, the column gender has a lot of **missing values**. Users who did not churn seem more likely to leave their gender empty. On the other hand, without considering missing values, female users churned slightly more than male users, but the numbers are really similar.

-* Note: As mentioned above in section "II. Data Description", Kaggle did not provide what each City number actually refers to



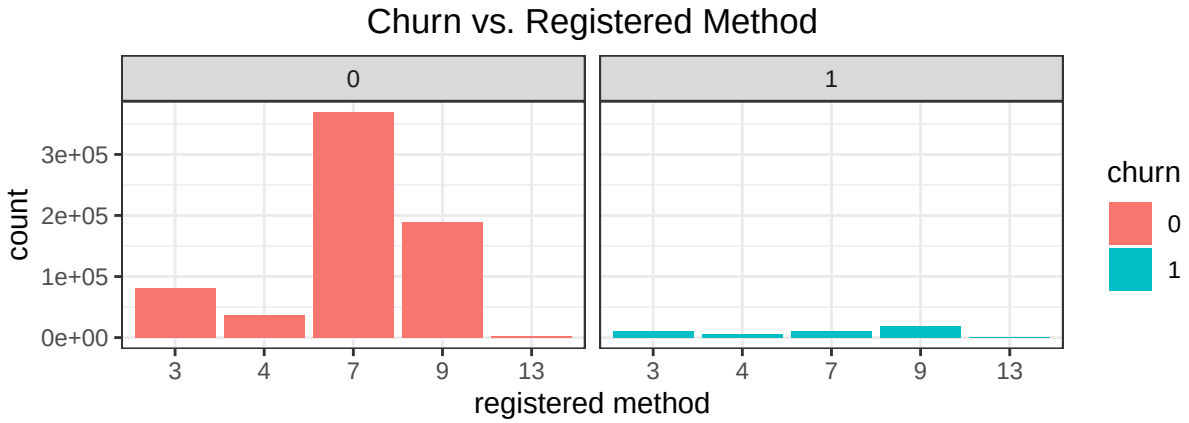


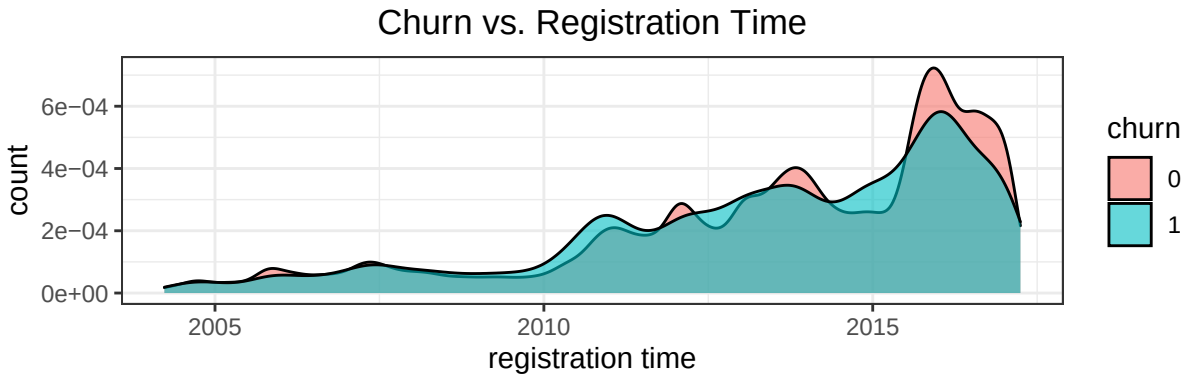
The box plots below show the distribution of the users' ages. We find that the column bd (age) has a lot of unrealistic **outliers**, ranging from -7000 to 2015. If we exclude the outliers, the ages of most users are between 22 and 30. Additionally, the median age of users who did not churn is slightly older than that of users who did churn.



The bar plot and line plot below indicate registration information**. Users who did not churn preferred to use method 7 and method 9 to register as a member. Additionally, the number of subscriptions has increased significantly since 2010, and the "non-churn rate" has been getting slightly higher since 2015.

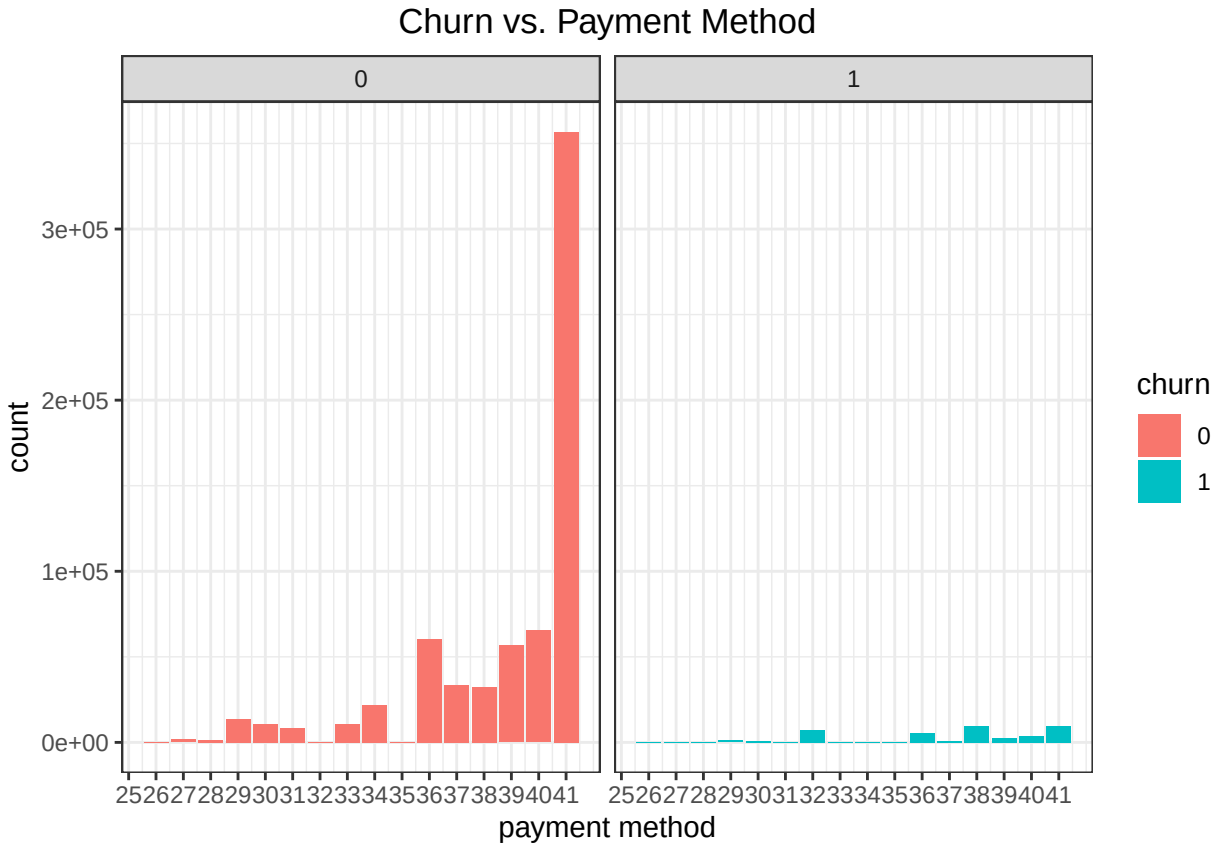
-** Note: As mentioned above in section "II. Data Description", Kaggle did not provide what each Registration Method number actually refers to





The bar plot below shows most users use method 36 to method 41 for payment***. Users who did not churn were more likely to use method 36, method 39, method 40, and method 41; users who churn preferred to use method 32, method 38 ,and method 41.

-*** Note: As mentioned above in section "II. Data Description", Kaggle did not provide what each Payment Method id number actually refers to



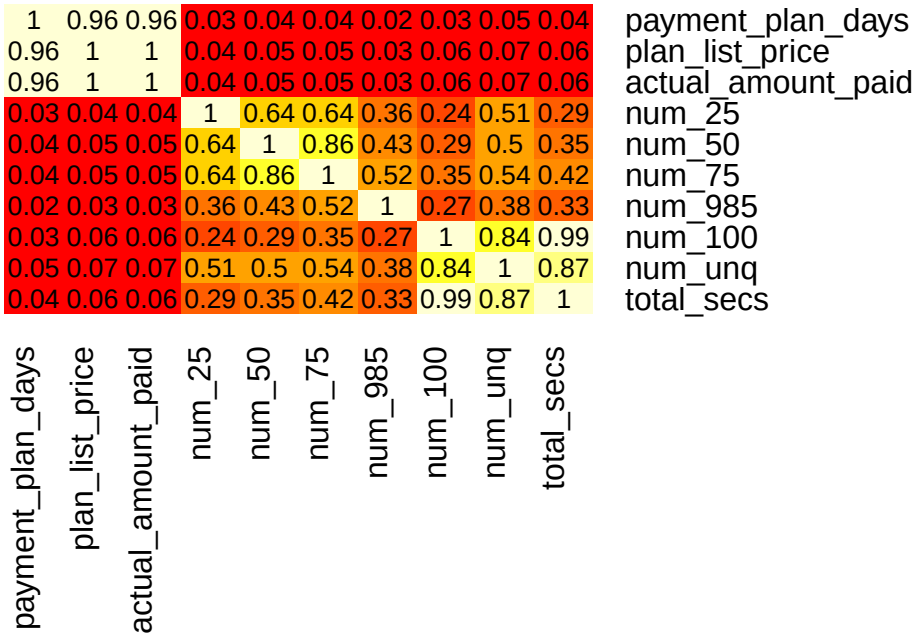
The table below compares the values for the mean amounts of instances in which churned customers have listened to portions of songs and the values for the mean amounts of instances in which non-churned customers have listened to portions of songs. This table illustrates that churned customers have a higher tendency to play songs for less than 25% of their original length compared to non-churn customers. However, churned customers have a higher tendency to play songs for longer amounts of time compared to non-churn customers. Additionally, churned customers have a higher tendency to play more unique songs and a higher tendency to further listen to songs for longer amounts of time compared to non-churn customers.

```
##   Churn   num_25   num_50   num_75  num_985  num_100  num_unq total_secs
## 1     0 109.3477 26.35549 16.64845 18.86472 542.6570 516.7876     141496
## 2     1 125.2071 30.54164 18.91417 21.04191 587.2263 573.6951     153375
```

From the correlation matrix below, we can find the variable num_50, num_75, num_100, num_unq, payment_plan_days, plan_list_price, and actual_amount_paid are highly correlated, implying this data set may have a **multicollinearity** problem.

```
##                     payment_plan_days plan_list_price actual_amount_paid num_25
## payment_plan_days                1.00            0.96               0.96   0.03
## plan_list_price                  0.96            1.00               1.00   0.04
## actual_amount_paid               0.96            1.00               1.00   0.04
## num_25                           0.03            0.04               0.04   1.00
## num_50                           0.04            0.05               0.05   0.64
## num_75                           0.04            0.05               0.05   0.64
## num_985                          0.02            0.03               0.03   0.36
## num_100                          0.03            0.06               0.06   0.24
## num_unq                          0.05            0.07               0.07   0.51
```

```
## total_secs                    0.04           0.06               0.06    0.29
##                     num_50 num_75 num_985 num_100 num_unq total_secs
## payment_plan_days     0.04   0.04    0.02    0.03    0.05       0.04
## plan_list_price       0.05   0.05    0.03    0.06    0.07       0.06
## actual_amount_paid    0.05   0.05    0.03    0.06    0.07       0.06
## num_25                0.64   0.64    0.36    0.24    0.51       0.29
## num_50                1.00   0.86    0.43    0.29    0.50       0.35
## num_75                0.86   1.00    0.52    0.35    0.54       0.42
## num_985               0.43   0.52    1.00    0.27    0.38       0.33
## num_100               0.29   0.35    0.27    1.00    0.84       0.99
## num_unq               0.50   0.54    0.38    0.84    1.00       0.87
## total_secs            0.35   0.42    0.33    0.99    0.87       1.00
```



V. Data Preparation & Model Building

Since this is a binary classification project (churn/non-churn), we used the classification algorithms of logistic regression, decision tree classification, and random forest classification to build models that could predict customer churn.

1 - Logistic regression

Data Preparation

The exploratory data analysis indicates that this data set may face the following problems: imbalanced data set, outliers, missing values, and multicollinearity. We deal with these challenges and other issues with the following changes.

First, we removed unneeded columns 'msno', 'transaction_date', 'membership_expire_date', and 'registration_init_time'.

Second, the feature bd indicates that user age has outliers as well as implausible values. The bd values range from as low as -7000 to as high as over 2015. However, age should range from 0 to 100 years in the real world. Thus, we remove the age values that are less than 0 or greater than 100. Additionally, we only keep the observations that fall between Q1 − (1.5)(IQR) and Q3 + (1.5)(IQR) and impute these removed values with the median value.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   28.00   28.00   28.39   28.00   49.00
```

Third, the feature gender has empty values for more than 60% of its values. However, we impute those empty values with "no" since we find users who did not churn are more likely to leave their gender empty from the exploratory data analysis.

Lastly, we convert the these four predictors 'city', 'gender', 'registered_via', 'payment_method_id into dummy variables and remove the original columns as well as one of the dummy variables from each categorical feature to prevent the issue of exact multicollinearity.

Logistic Regression Model

Our first algorithm is logistic regression. We split the data set in 80:20 ratio. Thus, 80% of the data will be used for the training set while the other 20% of the data will be used for the test set.

Before building a logistic regression model, we handle the imbalanced data set first. In our training data set, churn to non-churn ratio is 1 : 14. The imbalanced data set may result in poor predictive performance, especially for the minority class, so we solve this problem by resampling the training data set. Since there are many observations in our dataset, we choose to do the undersampling method.

```
##
##          0          1
## 0.93570201 0.06429799
```

```
##
##         0         1
## 0.5007957 0.4992043
```

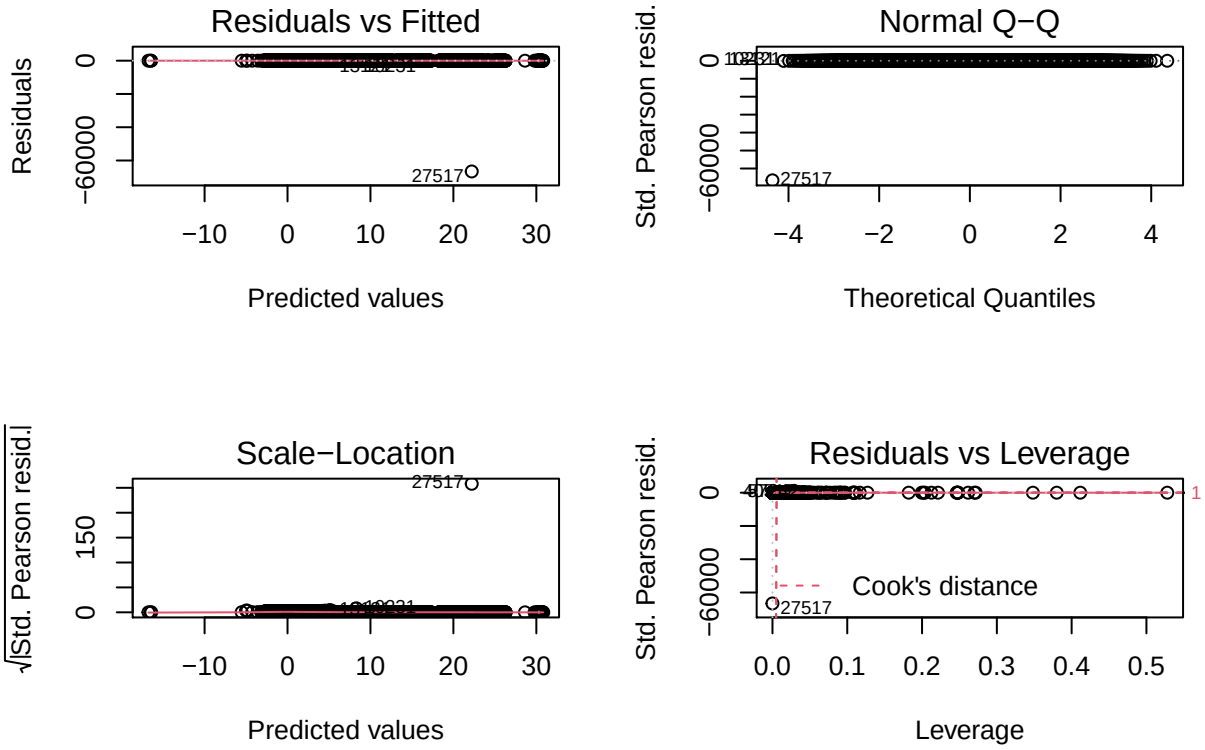
After solving the imbalanced data problem, we build a logistic regression model with all predictors. We find there is an outlier problem using diagnostic plots.

```
##
## Call:
## glm(formula = is_churn ~ ., family = "binomial", data = train.df_undersampling)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.6652  -0.7273  -0.3858   0.3824   3.1642
##
## Coefficients:
```

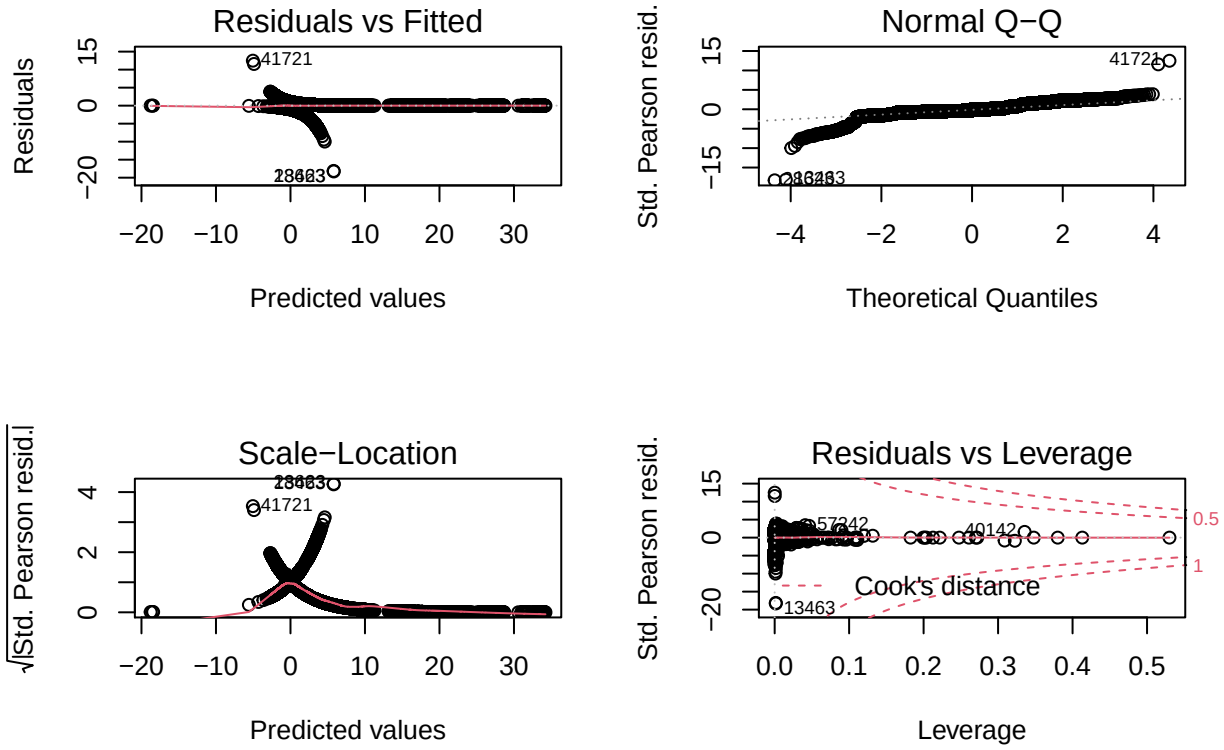
```
##                            Estimate        Std. Error z value
## (Intercept)            5.748358878726 1199.147894039518   0.005
## bd                     0.014236930763    0.002027165360   7.023
## num_25                -0.000024048669    0.000090870688  -0.265
## num_50                -0.001425785344    0.000515121784  -2.768
## num_75                 0.001622615472    0.000936317509   1.733
## num_985               -0.000709310575    0.000432506586  -1.640
## num_100               -0.000057721139    0.000080112496  -0.721
## num_unq                0.000024605972    0.000041967711   0.586
## total_secs             0.000000007259    0.000000364159   0.020
## payment_plan_days      0.005605283831    0.002474538443   2.265
## plan_list_price        0.031634094073    0.001022508769  30.938
## actual_amount_paid    -0.021645201457    0.000906130477 -23.888
## is_auto_renew         -1.591108298209    0.049029539611 -32.452
## is_cancel              4.203775014294    0.058601738529  71.735
## city_3                 0.504477414730    0.122978220826   4.102
## city_4                 0.468347474317    0.054317871112   8.622
## city_5                 0.410260340518    0.049454786307   8.296
## city_6                 0.342152432913    0.064699067448   5.288
## city_7                 0.526879776063    0.170991816834   3.081
## city_8                 0.497957503629    0.103025307735   4.833
## city_9                 0.459864620253    0.092185565811   4.988
## city_10                0.518019316523    0.113086304865   4.581
## city_11                0.371505802267    0.094991403302   3.911
## city_12                0.384280523308    0.091105493408   4.218
## city_13                0.423376664809    0.047097054634   8.989
## city_14                0.480842584125    0.069728735144   6.896
## city_15                0.427234690379    0.055741315766   7.665
## city_16                0.687094957146    0.287952940597   2.386
## city_17                0.158851824183    0.127713303655   1.244
## city_18                0.366693483568    0.107209870813   3.420
## city_19                0.040139867856    0.746297281956   0.054
## city_20                0.431435543878    0.353853534925   1.219
## city_21                0.490192744928    0.126058008741   3.889
## city_22                0.416202648045    0.056328692380   7.389
## gender_female          0.005315617337    0.027100535064   0.196
## gender_no             -0.126492623912    0.039189915353  -3.228
## registered_via_4       0.398488938975    0.042069025268   9.472
## registered_via_7      -0.088441117901    0.073661456487  -1.201
## registered_via_9      -0.055887852768    0.031482355491  -1.775
## registered_via_13     -0.138137409476    0.191255358033  -0.722
## payment_method_id_6   -8.395110251451 1807.967487817709  -0.005
## payment_method_id_8   -0.459507031965 1467.827169290173   0.000
## payment_method_id_10  -6.597172496120 1199.148278005972  -0.006
## payment_method_id_11 -23.140588265808 1830.558195069845  -0.013
## payment_method_id_12   3.759405668540 1207.102410545685   0.003
## payment_method_id_13  -0.830351308464 1206.412229947237  -0.001
## payment_method_id_14  -8.389140912271 1199.147980588527  -0.007
## payment_method_id_15  -5.810633692240 1199.147878495102  -0.005
## payment_method_id_16  -8.118537085812 1199.147913907884  -0.007
## payment_method_id_17  -1.618813642154 1199.147961713496  -0.001
## payment_method_id_18  -7.944066587613 1199.147975406348  -0.007
## payment_method_id_19  -7.885583483170 1199.147905993929  -0.007
## payment_method_id_20   4.428419658785 1202.514852757784   0.004
```

```
## payment_method_id_21   -7.421894330689 1199.147893993694  -0.006
## payment_method_id_22   -1.312285426106 1203.410325980111  -0.001
## payment_method_id_23   -8.409528758444 1199.147888047610  -0.007
## payment_method_id_26   10.259330062456 1248.366324272780   0.008
## payment_method_id_27   -8.490926887044 1199.147900341836  -0.007
## payment_method_id_28   -7.192748644391 1199.147869433821  -0.006
## payment_method_id_29   -8.166346832019 1199.147862895119  -0.007
## payment_method_id_30   -7.137601951032 1199.147868392903  -0.006
## payment_method_id_31   -8.985459887351 1199.147874063899  -0.007
## payment_method_id_32   -4.538033022396 1199.147931474326  -0.004
## payment_method_id_33   -8.464124423983 1199.147868641112  -0.007
## payment_method_id_34   -8.507735482461 1199.147866801510  -0.007
## payment_method_id_35   10.190343794885 1208.902804177165   0.008
## payment_method_id_36   -7.464276387069 1199.147861838827  -0.006
## payment_method_id_37   -8.388438875111 1199.147865740675  -0.007
## payment_method_id_38   -7.223092656576 1199.147863405140  -0.006
## payment_method_id_39   -7.222639715749 1199.147864679066  -0.006
## payment_method_id_40   -7.425352046220 1199.147864628372  -0.006
## payment_method_id_41   -7.119378436060 1199.147868520105  -0.006
##                                 Pr(>|z|)
## (Intercept)                     0.996175
## bd                     0.0000000000021704 ***
## num_25                          0.791281
## num_50                          0.005643 **
## num_75                          0.083100 .
## num_985                         0.101005
## num_100                         0.471217
## num_unq                         0.557669
## total_secs                      0.984097
## payment_plan_days               0.023501 *
## plan_list_price      < 0.0000000000000002 ***
## actual_amount_paid   < 0.0000000000000002 ***
## is_auto_renew        < 0.0000000000000002 ***
## is_cancel            < 0.0000000000000002 ***
## city_3                 0.0000409295574728 ***
## city_4               < 0.0000000000000002 ***
## city_5               < 0.0000000000000002 ***
## city_6                 0.0000001234129690 ***
## city_7                          0.002061 **
## city_8                 0.0000013425357823 ***
## city_9                 0.0000006086041644 ***
## city_10                0.0000046332658683 ***
## city_11                0.0000919371164259 ***
## city_12                0.0000246508184469 ***
## city_13              < 0.0000000000000002 ***
## city_14                0.0000000000053524 ***
## city_15                0.0000000000000179 ***
## city_16                         0.017026 *
## city_17                         0.213567
## city_18                         0.000625 ***
## city_19                         0.957106
## city_20                         0.222750
## city_21                         0.000101 ***
## city_22                0.0000000000001481 ***
```

```
## gender_female                    0.844497
## gender_no                        0.001248 **
## registered_via_4     < 0.0000000000000002 ***
## registered_via_7                 0.229890
## registered_via_9                 0.075863 .
## registered_via_13                0.470130
## payment_method_id_6              0.996295
## payment_method_id_8              0.999750
## payment_method_id_10             0.995610
## payment_method_id_11             0.989914
## payment_method_id_12             0.997515
## payment_method_id_13             0.999451
## payment_method_id_14             0.994418
## payment_method_id_15             0.996134
## payment_method_id_16             0.994598
## payment_method_id_17             0.998923
## payment_method_id_18             0.994714
## payment_method_id_19             0.994753
## payment_method_id_20             0.997062
## payment_method_id_21             0.995062
## payment_method_id_22             0.999130
## payment_method_id_23             0.994405
## payment_method_id_26             0.993443
## payment_method_id_27             0.994350
## payment_method_id_28             0.995214
## payment_method_id_29             0.994566
## payment_method_id_30             0.995251
## payment_method_id_31             0.994021
## payment_method_id_32             0.996981
## payment_method_id_33             0.994368
## payment_method_id_34             0.994339
## payment_method_id_35             0.993274
## payment_method_id_36             0.995033
## payment_method_id_37             0.994419
## payment_method_id_38             0.995194
## payment_method_id_39             0.995194
## payment_method_id_40             0.995059
## payment_method_id_41             0.995263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 103666  on 74778  degrees of freedom
## Residual deviance:  61758  on 74707  degrees of freedom
## AIC: 61902
##
## Number of Fisher Scoring iterations: 15
```



After removing outliers based on the cook's distance, there is no studentized residual falling outside the red



dotted lines.

Second, we test for multicollinearity with Variance Inflation Factors (VIF). Since a VIF exceeding five indicates a high correlation that may be problematic, we drop those predictors with VIF values higher than 5. The table below shows that the VIF values of left preditors are no larger than five.

```
##                  bd              num_25              num_50
##               FALSE               FALSE               FALSE
##              num_75             num_985             num_100
##                TRUE               FALSE                TRUE
##             num_unq          total_secs    payment_plan_days
##                TRUE                TRUE               FALSE
##     plan_list_price   actual_amount_paid        is_auto_renew
##                TRUE                TRUE               FALSE
##           is_cancel              city_3              city_4
##               FALSE               FALSE               FALSE
##              city_5              city_6              city_7
##               FALSE               FALSE               FALSE
##              city_8              city_9             city_10
##               FALSE               FALSE               FALSE
##             city_11             city_12             city_13
##               FALSE               FALSE               FALSE
##             city_14             city_15             city_16
##               FALSE               FALSE               FALSE
##             city_17             city_18             city_19
##               FALSE               FALSE               FALSE
##             city_20             city_21             city_22
##               FALSE               FALSE               FALSE
```

```
##        gender_female             gender_no     registered_via_4
##                FALSE                 FALSE                FALSE
##     registered_via_7     registered_via_9    registered_via_13
##                 TRUE                 FALSE                FALSE
##  payment_method_id_6  payment_method_id_8 payment_method_id_10
##                FALSE                 FALSE                 TRUE
## payment_method_id_11 payment_method_id_12 payment_method_id_13
##                FALSE                  TRUE                 TRUE
## payment_method_id_14 payment_method_id_15 payment_method_id_16
##                 TRUE                  TRUE                 TRUE
## payment_method_id_17 payment_method_id_18 payment_method_id_19
##                 TRUE                  TRUE                 TRUE
## payment_method_id_20 payment_method_id_21 payment_method_id_22
##                 TRUE                  TRUE                 TRUE
## payment_method_id_23 payment_method_id_26 payment_method_id_27
##                 TRUE                  TRUE                 TRUE
## payment_method_id_28 payment_method_id_29 payment_method_id_30
##                 TRUE                  TRUE                 TRUE
## payment_method_id_31 payment_method_id_32 payment_method_id_33
##                 TRUE                  TRUE                 TRUE
## payment_method_id_34 payment_method_id_35 payment_method_id_36
##                 TRUE                  TRUE                 TRUE
## payment_method_id_37 payment_method_id_38 payment_method_id_39
##                 TRUE                  TRUE                 TRUE
## payment_method_id_40 payment_method_id_41
##                 TRUE                  TRUE


##
## Call:
## glm(formula = is_churn ~ bd + num_25 + num_50 + num_75 + num_985 +
##     payment_plan_days + is_auto_renew + is_cancel + city_3 +
##     city_4 + city_5 + city_6 + city_7 + city_8 + city_9 + city_10 +
##     city_11 + city_12 + city_13 + city_14 + city_15 + city_16 +
##     city_17 + city_18 + city_19 + city_20 + city_21 + city_22 +
##     gender_female + gender_no + registered_via_4 + registered_via_9 +
##     registered_via_13 + payment_method_id_6 + payment_method_id_8 +
##     payment_method_id_11, family = "binomial", data = train.df_undersampling_nooutliers)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9348  -0.7831  -0.5809   0.3670   2.4585
##
## Coefficients:
##                          Estimate   Std. Error z value             Pr(>|z|)
## (Intercept)          -0.96068752   0.08450869 -11.368 < 0.0000000000000002 ***
## bd                    0.00319678   0.00192544   1.660               0.0969 .
## num_25               -0.00004140   0.00007936  -0.522               0.6019
## num_50               -0.00102125   0.00049556  -2.061               0.0393 *
## num_75                0.00118071   0.00090750   1.301               0.1932
## num_985              -0.00101140   0.00042789  -2.364               0.0181 *
## payment_plan_days     0.04648788   0.00159845  29.083 < 0.0000000000000002 ***
## is_auto_renew        -1.87549297   0.02675717 -70.093 < 0.0000000000000002 ***
## is_cancel             4.09517840   0.05789634  70.733 < 0.0000000000000002 ***
## city_3                0.53079953   0.11638198   4.561   0.0000050949484838 ***
```

```
## city_4                 0.53419477   0.05109452  10.455 < 0.0000000000000002 ***
## city_5                 0.51505261   0.04654151  11.067 < 0.0000000000000002 ***
## city_6                 0.46882329   0.06121612   7.658   0.0000000000000188 ***
## city_7                 0.69139169   0.15924049   4.342   0.0000141314724999 ***
## city_8                 0.58013476   0.09835178   5.899   0.0000000036666803 ***
## city_9                 0.53736611   0.08872985   6.056   0.0000000013937105 ***
## city_10                0.62782563   0.10839815   5.792   0.0000000069615996 ***
## city_11                0.47325622   0.09188888   5.150   0.0000002600564326 ***
## city_12                0.44015703   0.08611488   5.111   0.0000003199875372 ***
## city_13                0.49503124   0.04348242  11.385 < 0.0000000000000002 ***
## city_14                0.61224316   0.06633037   9.230 < 0.0000000000000002 ***
## city_15                0.54114763   0.05220899  10.365 < 0.0000000000000002 ***
## city_16                0.65720816   0.27475929   2.392               0.0168 *
## city_17                0.21298961   0.12254888   1.738               0.0822 .
## city_18                0.43554418   0.10351355   4.208   0.0000258091191575 ***
## city_19               -0.00339857   0.68121200  -0.005               0.9960
## city_20                0.45016925   0.34048821   1.322               0.1861
## city_21                0.60425405   0.11962811   5.051   0.0000004392633115 ***
## city_22                0.52850655   0.05278277  10.013 < 0.0000000000000002 ***
## gender_female          0.03190535   0.02615579   1.220               0.2225
## gender_no             -0.06689118   0.03774351  -1.772               0.0764 .
## registered_via_4       0.44841058   0.03612293  12.413 < 0.0000000000000002 ***
## registered_via_9      -0.09839998   0.02433284  -4.044   0.0000525657408095 ***
## registered_via_13      0.27486799   0.15693168   1.752               0.0799 .
## payment_method_id_6   -6.51634740 113.27979426  -0.058               0.9541
## payment_method_id_8    2.43673212  68.48984865   0.036               0.9716
## payment_method_id_11 -10.61156616 113.66462513  -0.093               0.9256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 103643  on 74762  degrees of freedom
## Residual deviance:  64839  on 74726  degrees of freedom
## AIC: 64913
##
## Number of Fisher Scoring iterations: 10

##                   bd              num_25              num_50
##             1.125809            1.928217            4.052085
##               num_75             num_985    payment_plan_days
##             4.939667            2.008941            1.013726
##        is_auto_renew           is_cancel              city_3
##             1.216723            1.014363            1.092326
##               city_4              city_5              city_6
##             1.746028            2.091104            1.442833
##               city_7              city_8              city_9
##             1.045655            1.127877            1.155148
##              city_10             city_11             city_12
##             1.107318            1.158888            1.191841
##              city_13             city_14             city_15
##             2.351892            1.350380            1.710646
##              city_16             city_17             city_18
##             1.016671            1.082907            1.119240
```

```
##              city_19              city_20              city_21
##             1.003151             1.011096             1.083031
##              city_22        gender_female            gender_no
##             1.683046             1.419874             3.750759
##     registered_via_4     registered_via_9    registered_via_13
##             1.139738             1.381408             1.006292
##  payment_method_id_6  payment_method_id_8 payment_method_id_11
##             1.000027             1.000015             1.000000

##                   bd               num_25               num_50
##                FALSE                FALSE                FALSE
##               num_75              num_985    payment_plan_days
##                FALSE                FALSE                FALSE
##        is_auto_renew            is_cancel               city_3
##                FALSE                FALSE                FALSE
##               city_4               city_5               city_6
##                FALSE                FALSE                FALSE
##               city_7               city_8               city_9
##                FALSE                FALSE                FALSE
##              city_10              city_11              city_12
##                FALSE                FALSE                FALSE
##              city_13              city_14              city_15
##                FALSE                FALSE                FALSE
##              city_16              city_17              city_18
##                FALSE                FALSE                FALSE
##              city_19              city_20              city_21
##                FALSE                FALSE                FALSE
##              city_22        gender_female            gender_no
##                FALSE                FALSE                FALSE
##     registered_via_4     registered_via_9    registered_via_13
##                FALSE                FALSE                FALSE
##  payment_method_id_6  payment_method_id_8 payment_method_id_11
##                FALSE                FALSE                FALSE
```

Third, we plan to use the backward elimination method to select the best performing model by removing those features that do not have a significant effect on the dependent variable. After optimizing the model with the backward elimination method, the AIC value reduces from 64913 to 64904. The predictors "payment_method_id_6", "payment_method_id_8", "city_20", "num_75", "city_19", "num_25", "num_50", and "gender_female" are removed. After the previous process, we get the best performing model as we can see below.

```
## Start:  AIC=64913.43
## is_churn ~ bd + num_25 + num_50 + num_75 + num_985 + payment_plan_days +
##     is_auto_renew + is_cancel + city_3 + city_4 + city_5 + city_6 +
##     city_7 + city_8 + city_9 + city_10 + city_11 + city_12 +
##     city_13 + city_14 + city_15 + city_16 + city_17 + city_18 +
##     city_19 + city_20 + city_21 + city_22 + gender_female + gender_no +
##     registered_via_4 + registered_via_9 + registered_via_13 +
##     payment_method_id_6 + payment_method_id_8 + payment_method_id_11
##
##                        Df Deviance   AIC
## - payment_method_id_6   1    64839 64911
## - city_19               1    64839 64911
```

```
## - payment_method_id_8   1    64839 64911
## - num_25                1    64840 64912
## - gender_female         1    64841 64913
## - city_20               1    64841 64913
## - num_75                1    64841 64913
## - payment_method_id_11  1    64841 64913
## <none>                       64839 64913
## - bd                    1    64842 64914
## - city_17               1    64842 64914
## - registered_via_13     1    64842 64914
## - gender_no             1    64843 64915
## - num_50                1    64844 64916
## - city_16               1    64845 64917
## - num_985               1    64846 64918
## - registered_via_9      1    64856 64928
## - city_18               1    64857 64929
## - city_7                1    64858 64930
## - city_3                1    64860 64932
## - city_21               1    64864 64936
## - city_12               1    64865 64937
## - city_11               1    64865 64937
## - city_10               1    64872 64944
## - city_8                1    64873 64945
## - city_9                1    64875 64947
## - city_6                1    64897 64969
## - city_14               1    64922 64994
## - city_22               1    64938 65010
## - city_15               1    64945 65017
## - city_4                1    64947 65019
## - city_5                1    64960 65032
## - city_13               1    64968 65040
## - registered_via_4      1    64993 65065
## - payment_plan_days     1    69966 70038
## - is_auto_renew         1    70180 70252
## - is_cancel             1    77771 77843
##
## Step:  AIC=64911.43
## is_churn ~ bd + num_25 + num_50 + num_75 + num_985 + payment_plan_days +
##     is_auto_renew + is_cancel + city_3 + city_4 + city_5 + city_6 +
##     city_7 + city_8 + city_9 + city_10 + city_11 + city_12 +
##     city_13 + city_14 + city_15 + city_16 + city_17 + city_18 +
##     city_19 + city_20 + city_21 + city_22 + gender_female + gender_no +
##     registered_via_4 + registered_via_9 + registered_via_13 +
##     payment_method_id_8 + payment_method_id_11
##
##                         Df Deviance   AIC
## - city_19                1    64839 64909
## - payment_method_id_8    1    64839 64909
## - num_25                 1    64840 64910
## - gender_female          1    64841 64911
## - city_20                1    64841 64911
## - num_75                 1    64841 64911
## - payment_method_id_11   1    64841 64911
## <none>                        64839 64911
```

```
## - bd                    1    64842 64912
## - city_17               1    64842 64912
## - registered_via_13     1    64842 64912
## - gender_no             1    64843 64913
## - num_50                1    64844 64914
## - city_16               1    64845 64915
## - num_985               1    64846 64916
## - registered_via_9      1    64856 64926
## - city_18               1    64857 64927
## - city_7                1    64858 64928
## - city_3                1    64860 64930
## - city_21               1    64864 64934
## - city_12               1    64865 64935
## - city_11               1    64865 64935
## - city_10               1    64872 64942
## - city_8                1    64873 64943
## - city_9                1    64875 64945
## - city_6                1    64897 64967
## - city_14               1    64922 64992
## - city_22               1    64938 65008
## - city_15               1    64945 65015
## - city_4                1    64947 65017
## - city_5                1    64960 65030
## - city_13               1    64968 65038
## - registered_via_4      1    64993 65063
## - payment_plan_days     1    69967 70037
## - is_auto_renew         1    70180 70250
## - is_cancel             1    77771 77841
##
## Step:  AIC=64909.43
## is_churn ~ bd + num_25 + num_50 + num_75 + num_985 + payment_plan_days +
##     is_auto_renew + is_cancel + city_3 + city_4 + city_5 + city_6 +
##     city_7 + city_8 + city_9 + city_10 + city_11 + city_12 +
##     city_13 + city_14 + city_15 + city_16 + city_17 + city_18 +
##     city_20 + city_21 + city_22 + gender_female + gender_no +
##     registered_via_4 + registered_via_9 + registered_via_13 +
##     payment_method_id_8 + payment_method_id_11
##
##                         Df Deviance   AIC
## - payment_method_id_8    1    64839 64907
## - num_25                 1    64840 64908
## - gender_female          1    64841 64909
## - city_20                1    64841 64909
## - num_75                 1    64841 64909
## - payment_method_id_11   1    64841 64909
## <none>                        64839 64909
## - bd                     1    64842 64910
## - city_17                1    64842 64910
## - registered_via_13      1    64842 64910
## - gender_no              1    64843 64911
## - num_50                 1    64844 64912
## - city_16                1    64845 64913
## - num_985                1    64846 64914
## - registered_via_9       1    64856 64924
```

```
## - city_18               1    64857 64925
## - city_7                1    64858 64926
## - city_3                1    64860 64928
## - city_21               1    64864 64932
## - city_12               1    64865 64933
## - city_11               1    64865 64933
## - city_10               1    64872 64940
## - city_8                1    64873 64941
## - city_9                1    64875 64943
## - city_6                1    64897 64965
## - city_14               1    64922 64990
## - city_22               1    64938 65006
## - city_15               1    64945 65013
## - city_4                1    64947 65015
## - city_5                1    64961 65029
## - city_13               1    64968 65036
## - registered_via_4      1    64993 65061
## - payment_plan_days     1    69967 70035
## - is_auto_renew         1    70181 70249
## - is_cancel             1    77771 77839
##
## Step:  AIC=64907.43
## is_churn ~ bd + num_25 + num_50 + num_75 + num_985 + payment_plan_days +
##     is_auto_renew + is_cancel + city_3 + city_4 + city_5 + city_6 +
##     city_7 + city_8 + city_9 + city_10 + city_11 + city_12 +
##     city_13 + city_14 + city_15 + city_16 + city_17 + city_18 +
##     city_20 + city_21 + city_22 + gender_female + gender_no +
##     registered_via_4 + registered_via_9 + registered_via_13 +
##     payment_method_id_11
##
##                         Df Deviance   AIC
## - num_25                 1    64840 64906
## - gender_female          1    64841 64907
## - city_20                1    64841 64907
## - num_75                 1    64841 64907
## - payment_method_id_11   1    64841 64907
## <none>                        64839 64907
## - bd                     1    64842 64908
## - city_17                1    64842 64908
## - registered_via_13      1    64842 64908
## - gender_no              1    64843 64909
## - num_50                 1    64844 64910
## - city_16                1    64845 64911
## - num_985                1    64846 64912
## - registered_via_9       1    64856 64922
## - city_18                1    64857 64923
## - city_7                 1    64858 64924
## - city_3                 1    64860 64926
## - city_21                1    64864 64930
## - city_12                1    64865 64931
## - city_11                1    64865 64931
## - city_10                1    64872 64938
## - city_8                 1    64873 64939
## - city_9                 1    64875 64941
```

```
## - city_6                1    64897 64963
## - city_14               1    64922 64988
## - city_22               1    64938 65004
## - city_15               1    64945 65011
## - city_4                1    64947 65013
## - city_5                1    64961 65027
## - city_13               1    64968 65034
## - registered_via_4      1    64993 65059
## - payment_plan_days     1    69969 70035
## - is_auto_renew         1    70181 70247
## - is_cancel             1    77771 77837
##
## Step:  AIC=64905.71
## is_churn ~ bd + num_50 + num_75 + num_985 + payment_plan_days +
##     is_auto_renew + is_cancel + city_3 + city_4 + city_5 + city_6 +
##     city_7 + city_8 + city_9 + city_10 + city_11 + city_12 +
##     city_13 + city_14 + city_15 + city_16 + city_17 + city_18 +
##     city_20 + city_21 + city_22 + gender_female + gender_no +
##     registered_via_4 + registered_via_9 + registered_via_13 +
##     payment_method_id_11
##
##                        Df Deviance   AIC
## - gender_female         1    64841 64905
## - num_75                1    64841 64905
## - city_20               1    64841 64905
## - payment_method_id_11  1    64842 64906
## <none>                       64840 64906
## - bd                    1    64843 64907
## - city_17               1    64843 64907
## - registered_via_13     1    64843 64907
## - gender_no             1    64843 64907
## - num_50                1    64845 64909
## - city_16               1    64845 64909
## - num_985               1    64847 64911
## - registered_via_9      1    64856 64920
## - city_18               1    64857 64921
## - city_7                1    64858 64922
## - city_3                1    64860 64924
## - city_21               1    64865 64929
## - city_12               1    64865 64929
## - city_11               1    64866 64930
## - city_10               1    64872 64936
## - city_8                1    64874 64938
## - city_9                1    64875 64939
## - city_6                1    64898 64962
## - city_14               1    64923 64987
## - city_22               1    64939 65003
## - city_15               1    64946 65010
## - city_4                1    64948 65012
## - city_5                1    64961 65025
## - city_13               1    64968 65032
## - registered_via_4      1    64993 65057
## - payment_plan_days     1    69971 70035
## - is_auto_renew         1    70187 70251
```

```
## - is_cancel             1    77787 77851
##
## Step:  AIC=64905.19
## is_churn ~ bd + num_50 + num_75 + num_985 + payment_plan_days +
##     is_auto_renew + is_cancel + city_3 + city_4 + city_5 + city_6 +
##     city_7 + city_8 + city_9 + city_10 + city_11 + city_12 +
##     city_13 + city_14 + city_15 + city_16 + city_17 + city_18 +
##     city_20 + city_21 + city_22 + gender_no + registered_via_4 +
##     registered_via_9 + registered_via_13 + payment_method_id_11
##
##                         Df Deviance   AIC
## - num_75                 1    64843 64905
## - city_20                1    64843 64905
## - payment_method_id_11   1    64843 64905
## <none>                        64841 64905
## - city_17                1    64844 64906
## - bd                     1    64844 64906
## - registered_via_13      1    64844 64906
## - num_50                 1    64846 64908
## - gender_no              1    64847 64909
## - city_16                1    64847 64909
## - num_985                1    64848 64910
## - registered_via_9       1    64857 64919
## - city_18                1    64858 64920
## - city_7                 1    64859 64921
## - city_3                 1    64861 64923
## - city_21                1    64866 64928
## - city_12                1    64866 64928
## - city_11                1    64867 64929
## - city_10                1    64873 64935
## - city_8                 1    64875 64937
## - city_9                 1    64877 64939
## - city_6                 1    64899 64961
## - city_14                1    64924 64986
## - city_22                1    64940 65002
## - city_15                1    64947 65009
## - city_4                 1    64949 65011
## - city_5                 1    64962 65024
## - city_13                1    64970 65032
## - registered_via_4       1    64994 65056
## - payment_plan_days      1    69974 70036
## - is_auto_renew          1    70189 70251
## - is_cancel              1    77788 77850
##
## Step:  AIC=64904.68
## is_churn ~ bd + num_50 + num_985 + payment_plan_days + is_auto_renew +
##     is_cancel + city_3 + city_4 + city_5 + city_6 + city_7 +
##     city_8 + city_9 + city_10 + city_11 + city_12 + city_13 +
##     city_14 + city_15 + city_16 + city_17 + city_18 + city_20 +
##     city_21 + city_22 + gender_no + registered_via_4 + registered_via_9 +
##     registered_via_13 + payment_method_id_11
##
##                         Df Deviance   AIC
## - city_20                1    64844 64904
```

```
## - payment_method_id_11  1    64845 64905
## <none>                       64843 64905
## - bd                    1    64846 64906
## - registered_via_13     1    64846 64906
## - city_17               1    64846 64906
## - num_50                1    64847 64907
## - gender_no             1    64848 64908
## - num_985               1    64848 64908
## - city_16               1    64848 64908
## - registered_via_9      1    64859 64919
## - city_18               1    64860 64920
## - city_7                1    64861 64921
## - city_3                1    64863 64923
## - city_21               1    64867 64927
## - city_12               1    64868 64928
## - city_11               1    64869 64929
## - city_10               1    64875 64935
## - city_8                1    64877 64937
## - city_9                1    64878 64938
## - city_6                1    64900 64960
## - city_14               1    64926 64986
## - city_22               1    64942 65002
## - city_15               1    64949 65009
## - city_4                1    64951 65011
## - city_5                1    64964 65024
## - city_13               1    64971 65031
## - registered_via_4      1    64996 65056
## - payment_plan_days     1    69974 70034
## - is_auto_renew         1    70193 70253
## - is_cancel             1    77790 77850
##
## Step:  AIC=64904.36
## is_churn ~ bd + num_50 + num_985 + payment_plan_days + is_auto_renew +
##     is_cancel + city_3 + city_4 + city_5 + city_6 + city_7 +
##     city_8 + city_9 + city_10 + city_11 + city_12 + city_13 +
##     city_14 + city_15 + city_16 + city_17 + city_18 + city_21 +
##     city_22 + gender_no + registered_via_4 + registered_via_9 +
##     registered_via_13 + payment_method_id_11
##
##                         Df Deviance   AIC
## - payment_method_id_11  1    64846 64904
## <none>                       64844 64904
## - city_17               1    64847 64905
## - bd                    1    64847 64905
## - registered_via_13     1    64847 64905
## - num_50                1    64849 64907
## - city_16               1    64850 64908
## - num_985               1    64850 64908
## - gender_no             1    64850 64908
## - registered_via_9      1    64860 64918
## - city_18               1    64861 64919
## - city_7                1    64862 64920
## - city_3                1    64864 64922
## - city_21               1    64869 64927
```

```
## - city_12              1    64869 64927
## - city_11              1    64870 64928
## - city_10              1    64876 64934
## - city_8               1    64878 64936
## - city_9               1    64879 64937
## - city_6               1    64901 64959
## - city_14              1    64926 64984
## - city_22              1    64942 65000
## - city_15              1    64949 65007
## - city_4               1    64951 65009
## - city_5               1    64964 65022
## - city_13              1    64971 65029
## - registered_via_4     1    64998 65056
## - payment_plan_days    1    69978 70036
## - is_auto_renew        1    70193 70251
## - is_cancel            1    77790 77848
##
## Step:  AIC=64904.31
## is_churn ~ bd + num_50 + num_985 + payment_plan_days + is_auto_renew +
##     is_cancel + city_3 + city_4 + city_5 + city_6 + city_7 +
##     city_8 + city_9 + city_10 + city_11 + city_12 + city_13 +
##     city_14 + city_15 + city_16 + city_17 + city_18 + city_21 +
##     city_22 + gender_no + registered_via_4 + registered_via_9 +
##     registered_via_13
##
##                     Df Deviance   AIC
## <none>                   64846 64904
## - city_17            1    64849 64905
## - bd                 1    64849 64905
## - registered_via_13  1    64849 64905
## - num_50             1    64851 64907
## - city_16            1    64852 64908
## - num_985            1    64852 64908
## - gender_no          1    64852 64908
## - registered_via_9   1    64862 64918
## - city_18            1    64863 64919
## - city_7             1    64864 64920
## - city_3             1    64866 64922
## - city_21            1    64871 64927
## - city_12            1    64871 64927
## - city_11            1    64872 64928
## - city_10            1    64878 64934
## - city_8             1    64880 64936
## - city_9             1    64881 64937
## - city_6             1    64903 64959
## - city_14            1    64928 64984
## - city_22            1    64944 65000
## - city_15            1    64951 65007
## - city_4             1    64953 65009
## - city_5             1    64966 65022
## - city_13            1    64973 65029
## - registered_via_4   1    65000 65056
## - payment_plan_days  1    69980 70036
## - is_auto_renew      1    70196 70252
```

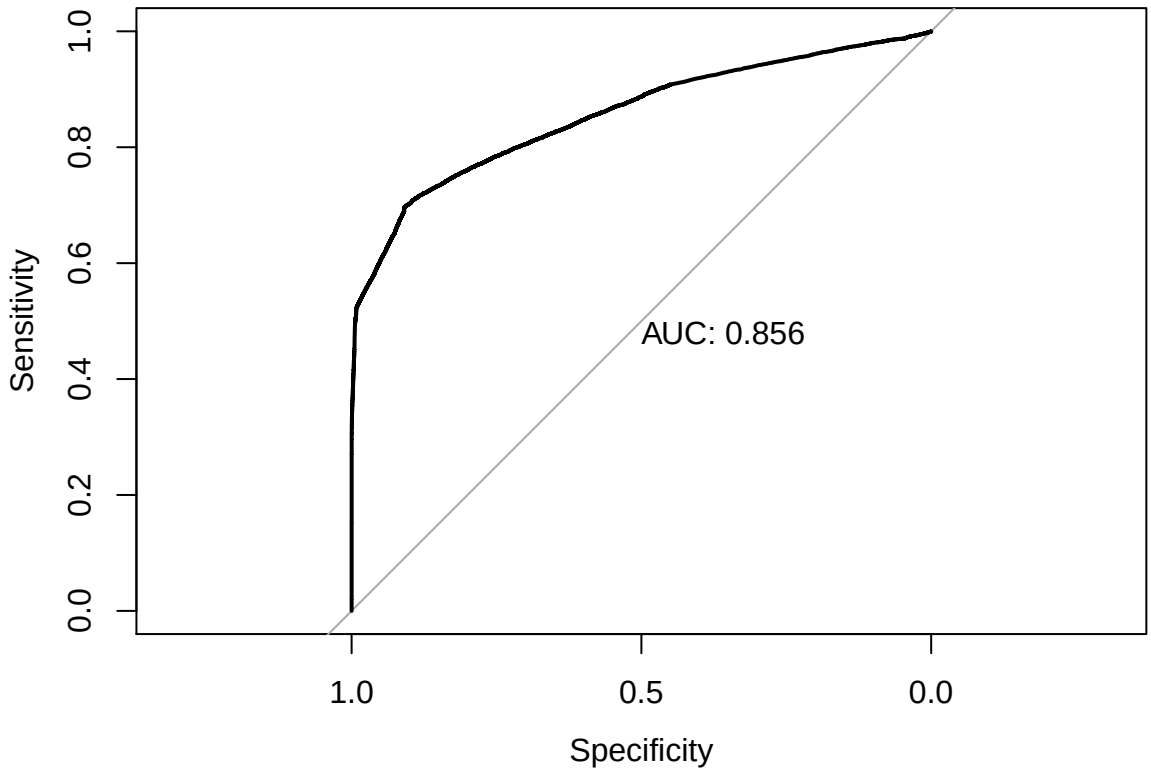
```
## - is_cancel          1    77793 77849
```

In the last step, we will measure if our customer churn predictive model is good. Below is the confusion matrix using a cutoff of 0.5 that we get from the final model. We mainly focus on sensitivity and F1_Score to measure our model performance because a highly sensitive model is useful to predict who will churn. This model achieves 0.6896 sensitivity will identify around 68.96% of churned customers but will miss around 31.04% of churned customers. Additionally, we want to send retention messages to retain all users who would like to churn and try not to bother those who do not plan to churn, so F1-score is a useful measure matrix in this case.

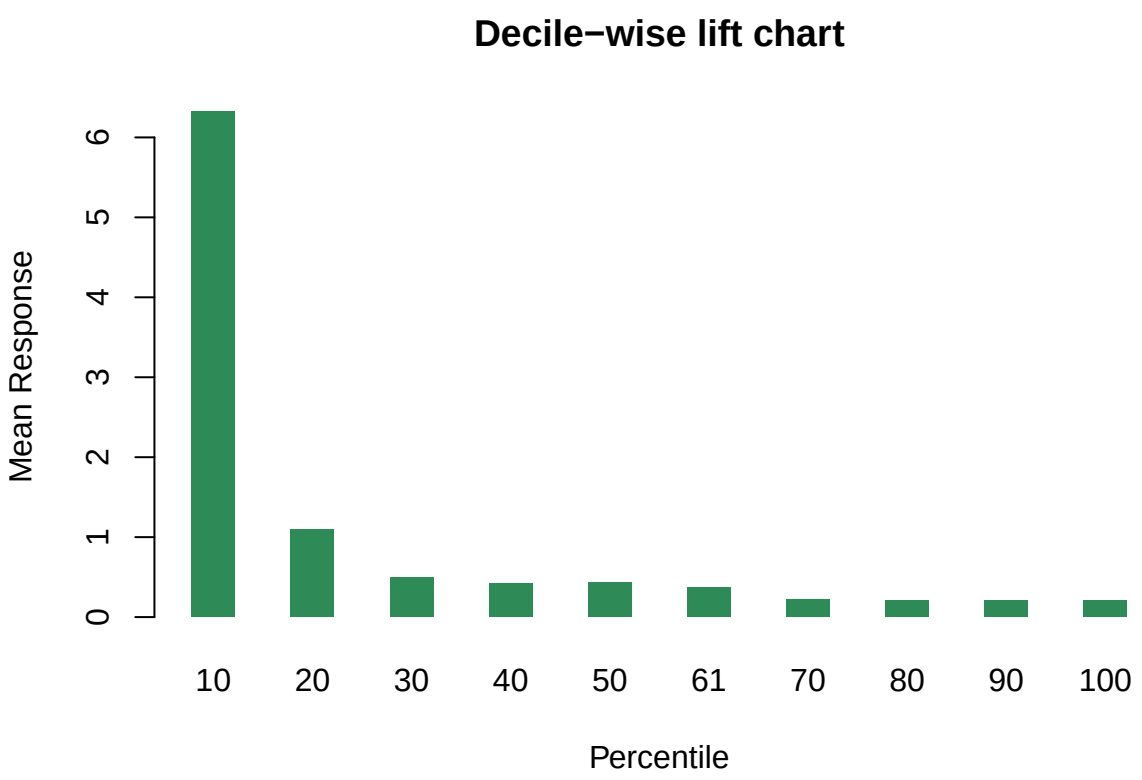
```
##        0    1
## 0 123536 2878
## 1  12335 6395
```

```
##             logistic_regression
## accuracy              0.8951868
## precision             0.3414309
## sensitivity           0.6896366
## specificity           0.9092154
## F1_Score              0.9419982
```

The area under the curve (AUC) tells how much the model can distinguish between classes. AUC of this model is 0.856.



From the decile-wise lift chart below, we can find the bars decreasing order from left to right, indicating this is a good prediction model.



2 - Decision Tree Classification

Data Preparation

Data preparation for the decision tree classification model is the same as before for the logistic regression model.

Decision Tree Classification Model

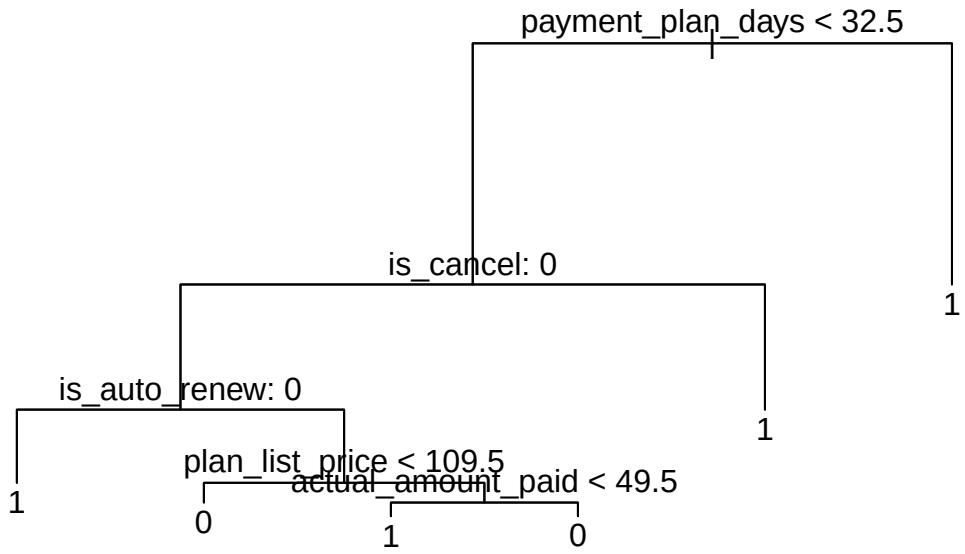
Our second algorithm is decision tree classification. We split the data set in 80:20 ratio. Thus, 80% of the data will be used for the training set while the other 20% of the data will be used for the test set.

We also handle the imbalanced data set by using the undersampling method for the decision tree classification model to let the churn to non-churn ratio be closer to 1 : 1 in the training set. As we can see below, the proportion of non-churned customers in the new training set is 0.5007957 while the proportion of churned customers in the new training set is 0.4992043.

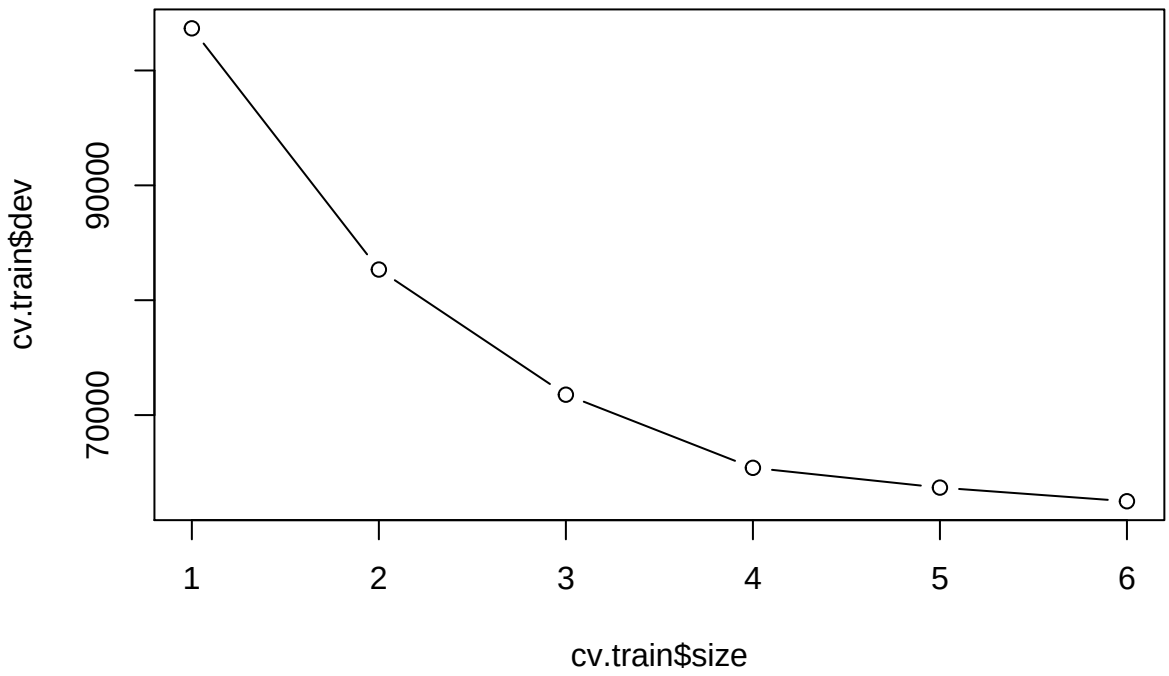
```
##
##         0         1
## 0.5007957 0.4992043
```

From this simple decision tree below, we can see that the five variables that are actually used in this simple tree's construction are: 'payment_plan_days', 'is_cancel', 'is_auto_renew', 'plan_list_price', and 'actual_amount_paid'. These five variables will be used when constructing the recursive partitioning decision tree later.

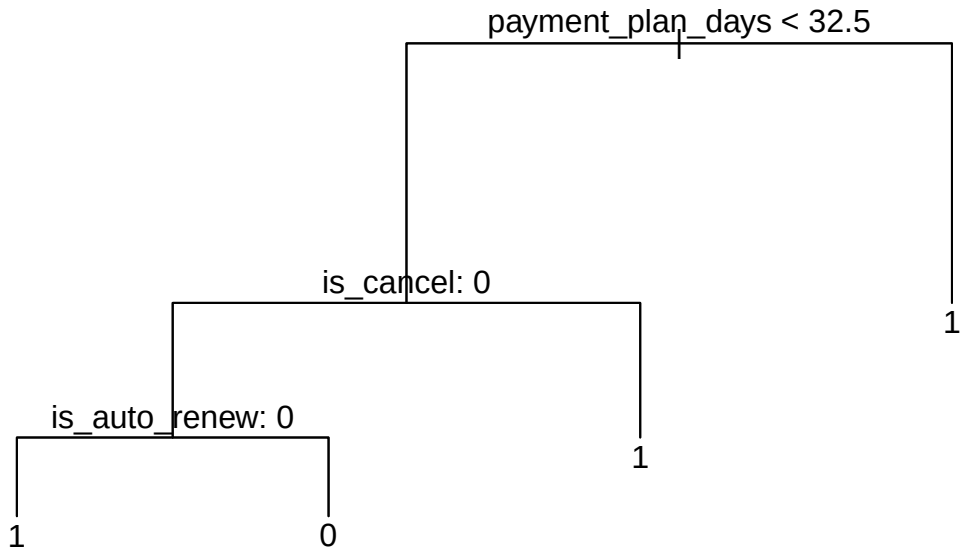
```
##
## Classification tree:
## tree(formula = is_churn ~ ., data = train.df_undersampling)
## Variables actually used in tree construction:
## [1] "payment_plan_days"  "is_cancel"          "is_auto_renew"
## [4] "plan_list_price"    "actual_amount_paid"
## Number of terminal nodes:  6
## Residual mean deviance:  0.8357426 = 62490.98 / 74773
## Misclassification error rate: 0.1870311 = 13986 / 74779
```



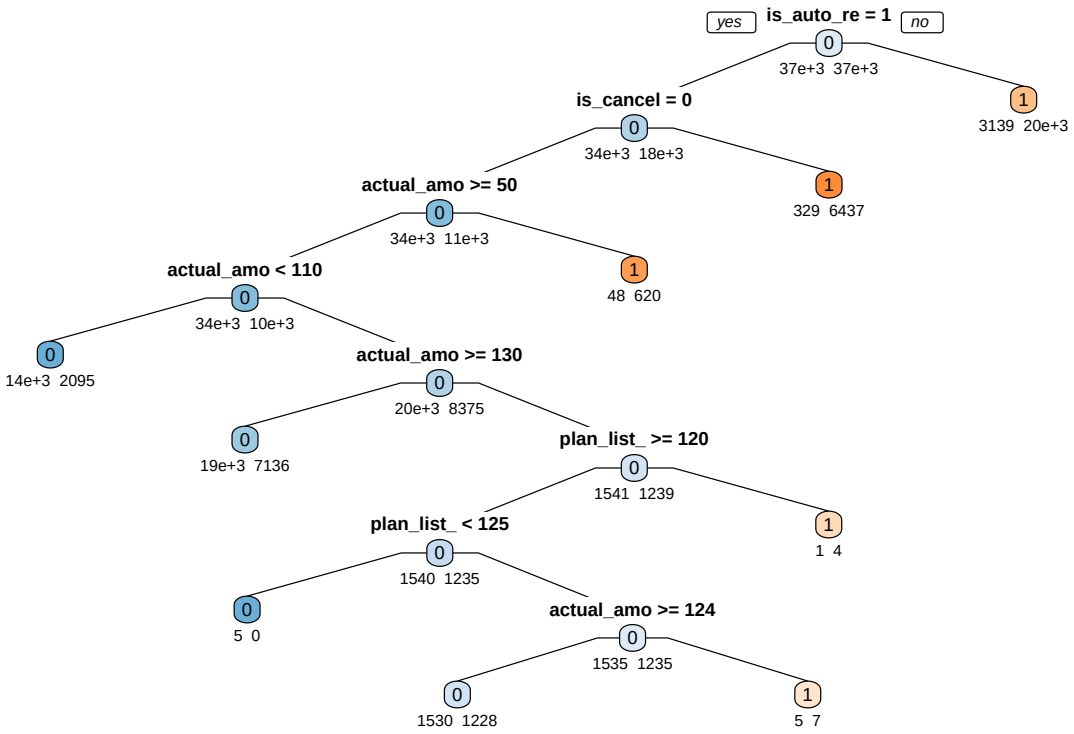
From the plot below, the optimal value for the number of terminal nodes should be 4.



Pruning the simple decision tree from above and setting the number of terminal nodes to be 4 because of the plot above. The resulting pruned simple decision is below with 4 terminal nodes.



Below is the recursive partitioning decision tree with the five variables of 'payment_plan_days', 'is_cancel', 'is_auto_renew', 'plan_list_price', and 'actual_amount_paid'. As mentioned above, these five variables were the only variables that were used in the construction of the simple decision tree above. Additionally, the rules for determining customer churn are below too.



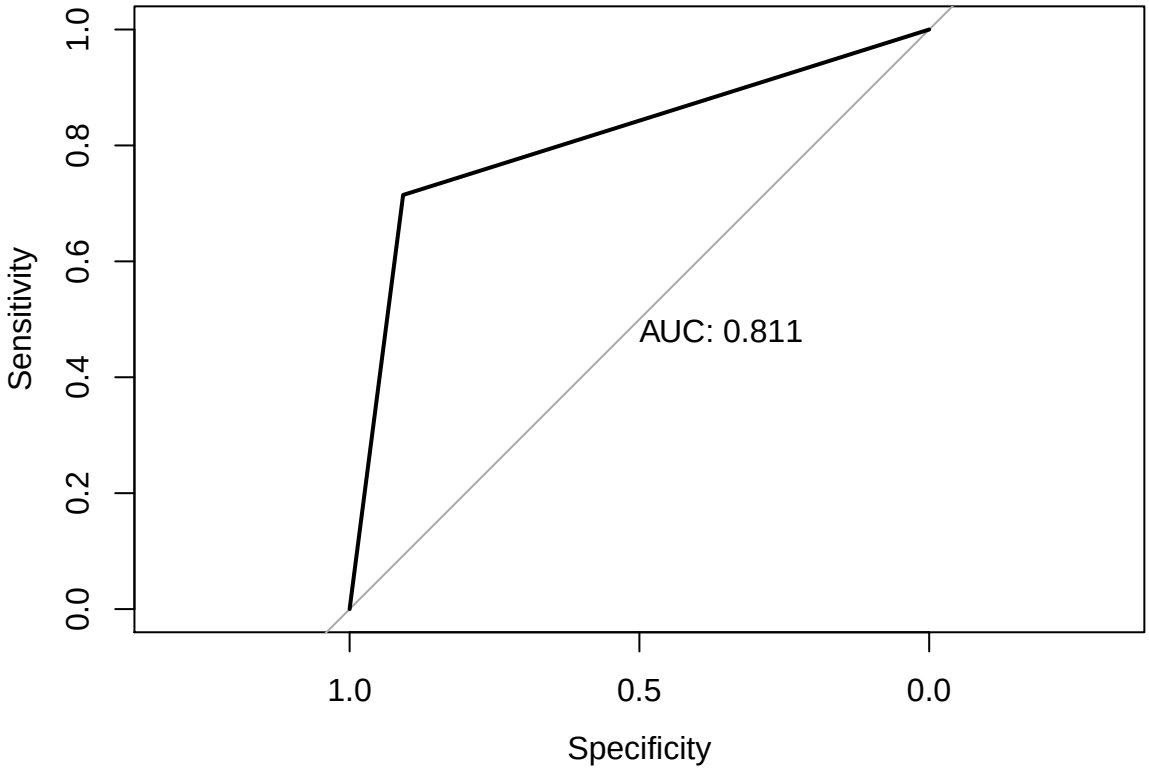

```
##  is_churn
##      0.00 when is_auto_renew is 1 & is_cancel is 0 & actual_amount_paid is 110 to 130 & plan_list_pr
##      0.13 when is_auto_renew is 1 & is_cancel is 0 & actual_amount_paid is  50 to 110
##      0.27 when is_auto_renew is 1 & is_cancel is 0 & actual_amount_paid >=        130
##      0.45 when is_auto_renew is 1 & is_cancel is 0 & actual_amount_paid is 124 to 130 & plan_list_pr
##      0.58 when is_auto_renew is 1 & is_cancel is 0 & actual_amount_paid is 110 to 124 & plan_list_pr
##      0.80 when is_auto_renew is 1 & is_cancel is 0 & actual_amount_paid is 110 to 130 & plan_list_pr
##      0.86 when is_auto_renew is 0
##      0.93 when is_auto_renew is 1 & is_cancel is 0 & actual_amount_paid <   50
##      0.95 when is_auto_renew is 1 & is_cancel is 1
```

Below is the confusion matrix for the recursive partitioning decision tree classification model. Additionally, below is also a table that describes the different metrics for the recursive partitioning decision tree classification model. This table depicts the metrics of accuracy, precision, sensitivity, specificity, and F1_Score.

```
##
## predictions      0      1
##           0 123328   2647
##           1  12543   6626
```

```
##             decision_tree
## accuracy        0.8953453
## precision       0.9361117
## sensitivity     0.7145476
## specificity     0.9076845
## F1_Score        0.9419888
```

Area under the curve (AUC) of the recursive partitioning decision tree classification model is 0.811.



3-Random Forest Classification

Data Preparation

Our last algorithm that we used is random forest classification. Data preparation for the random forest classification model is the same as before for the logistic regression model.

Random Forest Classification Model

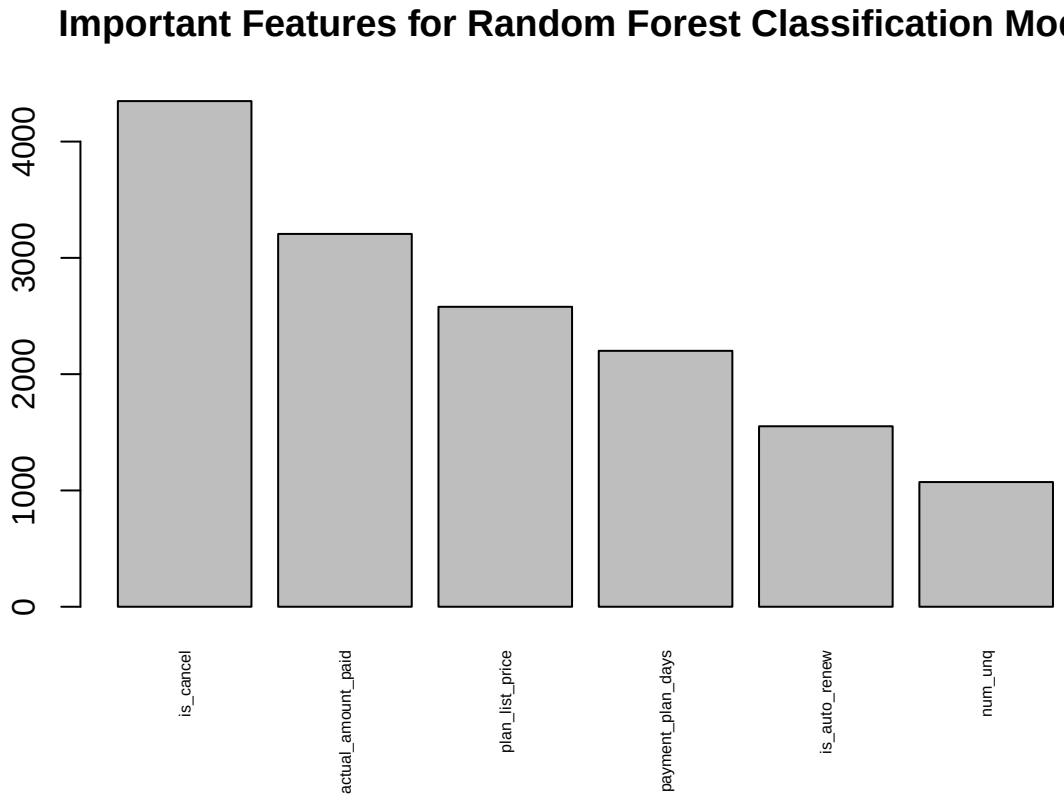
We split the data set in 80:20 ratio. Thus, 80% of the data will be used for the training set while the other 20% of the data will be used for the test set.

We also handle the imbalanced data set by using the undersampling method for the random forest classification model to let the churn to non-churn ratio be closer to $1:1$ in the training set. As we can see below, the proportion of non-churned customers in the new training set is 0.5007957 while the proportion of churned customers in the new training set is 0.4992043.

```
##
##         0         1
## 0.5007957 0.4992043
```

The top 6 variables that are the most important predictors for predicting customer churn in the Random Forest model, in order from most important to least important, are 'is_cancel', 'actual_amount_paid', 'plan_list_price', 'payment_plan_days', 'is_auto_renew', and 'num_unq'.

```
##          is_cancel actual_amount_paid    plan_list_price  payment_plan_days
##        4348.196970        3205.989391        2579.250424        2200.575217
##      is_auto_renew            num_unq
##        1551.860317        1072.207344
```

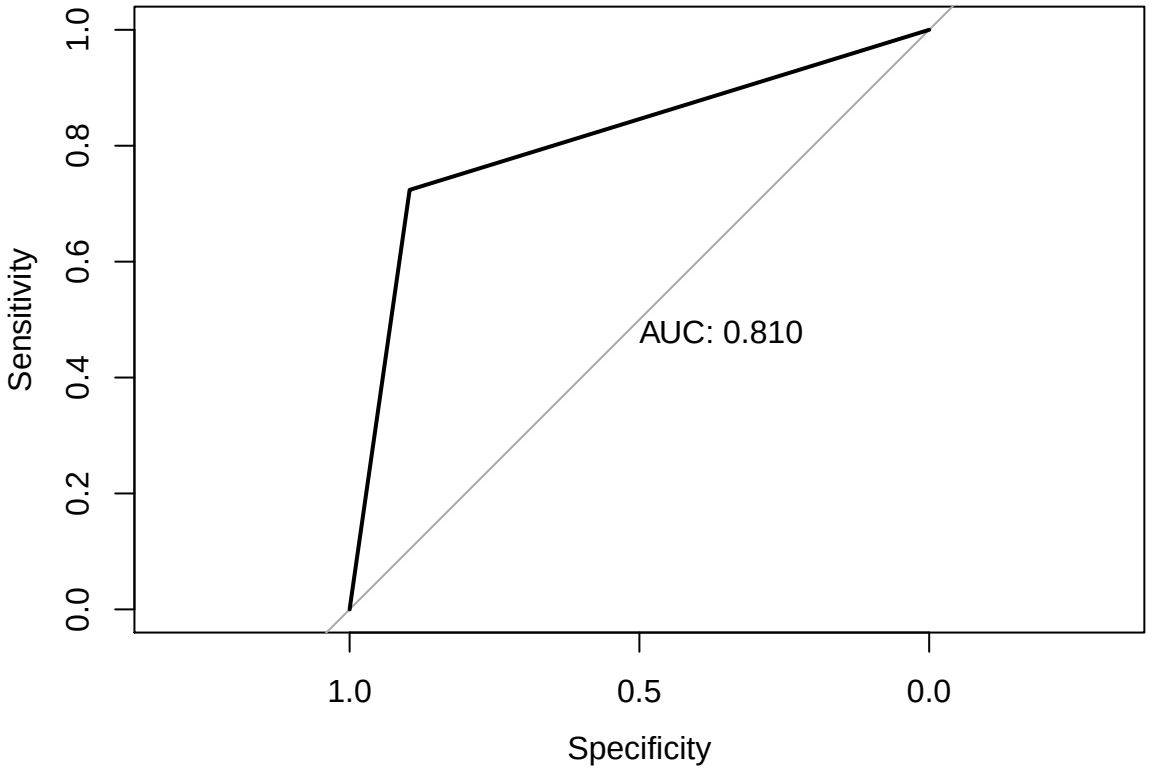


Below is the confusion matrix for the random forest classification model. Additionally, below is also a table that describes the different metrics for the random forest classification model. This table depicts the metrics of accuracy, precision, sensitivity, specificity, and F1_Score.

```
##
## predictions      0      1
##           0 121798   2561
##           1  14073   6712
```

```
##             random_forest
## accuracy     0.8853965717
## precision    0.9361117235
## sensitivity  0.7238218484
## specificity  0.8964238138
## F1_Score     0.9360796219
```

Area under the curve (AUC) of the random forest classification model is 0.810.



VI. Comparing Algorithms

Below is our performance comparison matrix from the three machine learning models that we built. We are choosing the best model based off the sensitivity, F1_Score, and AUC metrics. Since we want to identify all of the churned customers as precisely as possible, and the F1_Score and AUC of the three models are all good, we are mainly focusing on the sensitivity measure for our model performance because a highly sensitive model is useful to effectively identify and predict the customers who will churn. Therefore, we are choosing the random forest classification model as our best model because the random forest classification model had the highest sensitivity value among the 3 models with a sensitivity value of 0.724. The sensitivity value of 0.724 means this model will identify around 72% of churned customers but will miss around 28% of churned customers.

```
##             logistic_regression decision_tree random_forest
## accuracy                  0.895         0.895         0.885
## precision                 0.341         0.936         0.936
## sensitivity               0.690         0.715         0.724
## specificity               0.909         0.908         0.896
## F1_Score                  0.942         0.942         0.936
## auc                       0.856         0.811         0.810
```

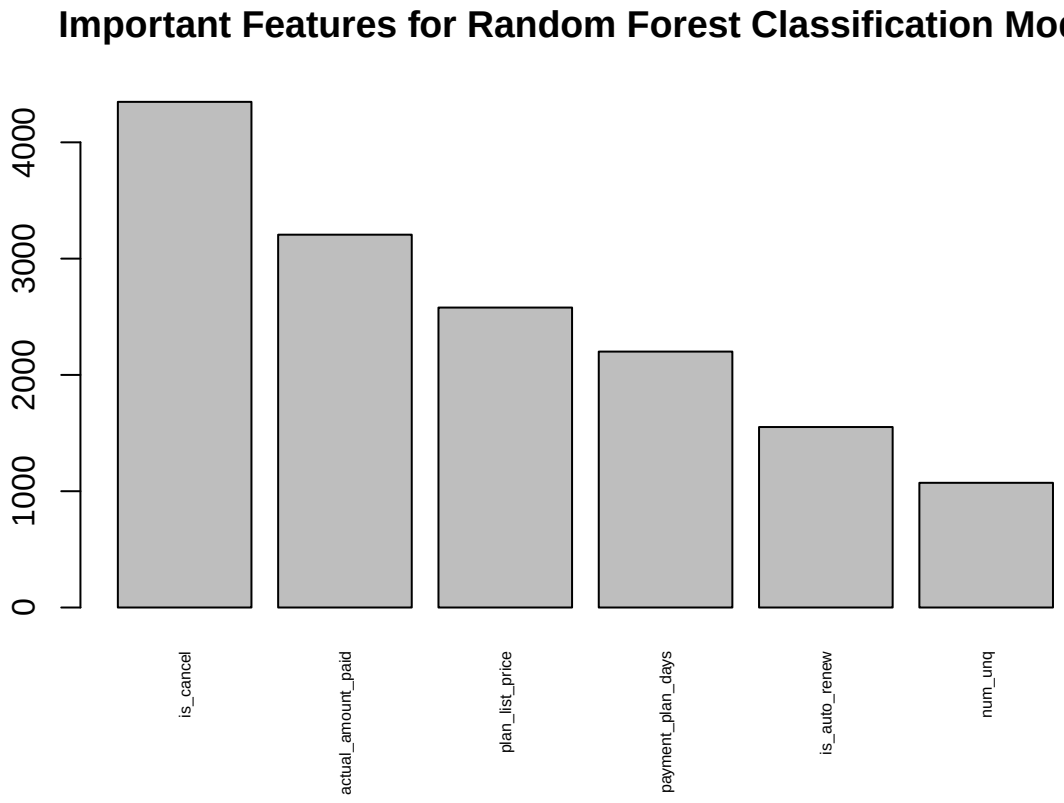
VII. Conclusion

From the results of the random forest classification model, the top five most influential variables for predicting customer churn are 'is_cancel', 'actual_amount_paid', 'plan_list_price', 'payment_plan_days', and

'is_auto_renew'. Below are five insights that we gained from this random forest classification model:

1. Customers who canceled their memberships in the transactions are more likely to churn than customers who did not cancel their memberships in the transactions.
2. Customers who have higher actual membership payments are more likely to churn than customers who have lower actual membership payments.
3. Customers who have higher membership plan payments are more likely to churn than customers who have lower membership plan payments.
4. Customers who have longer lengths of membership plans are more likely to churn than customers who have shorter lengths of membership plans.
5. Customers whose membership plans are auto-renew are more likely to churn than customers whose membership plans are not auto-renew.

The customers who resemble the criteria above are the ones who seem more likely to churn. Therefore, we highly recommend KKBox to develop some related marketing strategies based on these points listed above to retain the customers who seem more likely to churn. KKBox can use this analysis above to see which customers closely resemble these characteristics above to send retention messages, coupons, discounts, promo deals, etc. to these customers who fit the criteria above to try and convince/persuade these customers to continue their subscriptions and not churn.



VIII. References

Website

1. Datasets source from Kaggle:
   https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data

2. KKBox help center:
   https://help.kkbox.com/tw/en/billing/cancel-change/1338?p=kkbox