Ticket Redemption Data Analysis

Stat 427 Project

Wenke Huang (whuang67@illinois.edu)
Shuang Peng (speng9@illinois.edu)
Xinming Yang (xyang104@illinois.edu)
Yuanjing Zhu(yz10@illinois.edu)

Table of Contents

1. Introduction

The Krannert Center for the Performing Arts is an educational and performing arts complex on the campus of the University of Illinois at Urbana Champaign. Krannert Center have multiple theaters. The largest one is Foellinger Great Hall, with about 2,000 seats. It attracts world famous artists like Chicago Symphony Orchestra to perform every year.

The Krannert Center does not resell or oversell its tickets, so unredeemed tickets are wasted seats to the event. The redemption rates are typically 80% to 85%, even if all tickets have been distributed.

So, in this project, we aim to find out the reason why certain tickets are not redeemed and how it is related to the purchase behavior of customers. Based on that, we also want to detect some risky signal that indicate potential unredemption of the tickets.

Our dataset is ticket redemption data, which includes ticket information from July, 2016 to Feburary, 2017, mostly from fall semester, 2016. There are more than 50,000 observations and 20 variables in the original dataset. Each observation represents a ticket. The response variable Redeemed is a binary variable to indicate whether the ticket is redeemed or not.

In the following sections, we will first introduce the data processing steps we perform. And then, data visualization will be used to explore the relationships between the variables. After that, we perform statistical tests and build prediction models to further explore the data and statistical results will be presented. Finally, we will have some discussion on the following three questions of interest:

- What is the relationship between delivery option and redemption rates?
- What is the relationship between purchase behaviors and redemption rates?
- What are the risky signals that indicate potential unredemption?

2. Feature Engineering

In this section, we introduce how we perform data transformation, create new variables, impute missing values and remove outliers. After feature engineering, there are 13 main predictors to be further explored:

- Capacity
- Category
- DaysBetween
- Disposition
- EventPurchasedTotal
- Price
- PriceType
- Producer
- PurchasedOnline
- QuantityPurchased
- QuantityPurchasedTotal
- Theatre
- TicketSold

2.1 Summary of variables

Table 2.1 summarizes the information of selected variables before feature engineering.

Table 2.1 Summary of selected variables

Name	Type	Fact
PriceType	Factor	23 levels: 'SA','SA4','SAF','SAS', 'SC','SC4','SCF','SCS', 'SU','SU4','SUF','SUS', 'UI','UIC','UIF','UIS', 'YT','YT2','YTF','YTS', 'C', 'P', and 'SP'
CustomerName	Factor	Identification of customer
EventCode	Factor	Identification of event
DaysBetween	Numeric	The number of days between purchase date and event date. 0 means the ticket is purchased on the event date.
Theater	Factor	6 levels: the number behind each theater is the seat capacity of that theater Foellinger Great Hall (2066), Tryon Festival Theatre (979), Colwell Playhouse (641), Studio Theatre (200), FGH Stage Salon Style (161), Krannert Room
Category	Factor	12 levels
Capacity	Numeric	Sales rate.

		Definition: the amount of tickets sold divided by the capacity of tickets clients had to sell.
Producer	Factor	7 levels: Marquee, School of music, Illinois Theatre, Lyric Theatre at Illinois Dance at Illinois, Champaign Urbana Symphony Orchestra, Sunfonia daCamera
Disposition	Factor	6 levels: C(counter), M(mailed), MD1(mobile delivery pdf only), MD2(mobile delivery+pdf+passbook), PAH (print at home), W(will call)
Redeemed	Factor	2 levels: Yes/No Are coded into 1s and 0s, with 1 represents Yes and 0 represents No.

2.2 Data transformation

The categorical variable PriceType originally has 23 levels. We combine some levels into one level according their realistic meaning.

- 'SA','SA4','SAF', and 'SAS' are combined as 'SA'(Standard Admission)
- 'SC','SC4','SCF', and 'SCS' are combined as 'SC'(Senior Citizen)
- 'SU','SU4','SUF', and 'SUS' are combined as 'SU'(Non UI)
- 'UI','UIC','UIF', and 'UIS' are combined as 'UI'
- 'YT','YT2','YTF', and 'YTS' are combined as 'YT'(Youth)

After transformation, variable PriceType has 8 levels, 'SA', 'SC', 'SU', 'UI', 'YT', 'C', 'P', and 'SP'.

2.3 Creating New variables

To better analyze the data, we created 3 new variables.

- TicketSold:
  This new variable summarizes how many tickets are sold for each event. We created this variable mainly to impute values under '% of capacity sold'. We would explain how we used TicketSold to impute '% of capacity sold' later.

- QuantityPurchasedTotal:

This new variable summarizes how many tickets each customer purchased for the whole season.

- EventPurchasedTotal:
  We defined EventPurchasedTotal as: for each customer, how many different events did they purchase for the whole season.

Here is an example of how QuantityPurchasedTotal and EventPurchasedTotal work.

Table 2.2 Tickets purchased by customer 3 in 2016 Fall

Customer Number	Customer Name	Event Code	Quantity Purchased	Quantity Purhcased Total	Event Purchased Total
3	Prosise, Michael & Katherine	RUBB	2	9	5
3	Prosise, Michael & Katherine	RUBB	2	9	5
3	Prosise, Michael & Katherine	CUSB	2	9	5
3	Prosise, Michael & Katherine	925	1	9	5
3	Prosise, Michael & Katherine	ART	2	9	5
3	Prosise, Michael & Katherine	ART	2	9	5
3	Prosise, Michael & Katherine	CUSB	2	9	5
3	Prosise, Michael & Katherine	RU19	2	9	5
3	Prosise, Michael & Katherine	RU19	2	9	5

In this example, we included all information for customer 3, Prosise, Michael & Katherine. As we can see, for the whole season, they went to 5 different events. They are RUBB, CUSB, 925, ART and RU19. That is why, the EventPurchasedTotal equals 5.

The customer name showed up 9 times in the dataset and that is the Quantity Purchased Total. That is because each observation represents one piece of ticket information. We cannot simply add up Quantity Purchased to get QuantityPurchasedTotal, since some information is replicated.

2.4 Imputing missing values

The variable Capacity includes about 35% of missing data, so we could not use it directly in our analysis. But based on its importance, we don’t want to give up this variable.

The meaning of Capacity is the amount of tickets sold divided by the capacity of tickets clients had to sell. So we impute the missing values of Capacity as TicketSold divided by the available seats of the Theatre. We kept non-missing values as they were.

2.5 Removing outliers

During our analysis, we noticed 'Musicals' under variable 'Category' acted differently from other categories. They have 100% of Capacity Sold for every single show, but their redemption rates are 0s. The following table summarizes information of 'Musicals'.

Table 2.3 Summary of tickets of Category 'Musicals'

Category	Event Code	Performance Title	Redeemed	Count
Musicals	DR5	Dreamgirls	0	177
Musicals	DR6	Dreamgirls	0	175
Musicals	DR7	Dreamgirls	0	171

After removing these outliers, variable 'Category' only has 11 levels.

2.6 Correlation

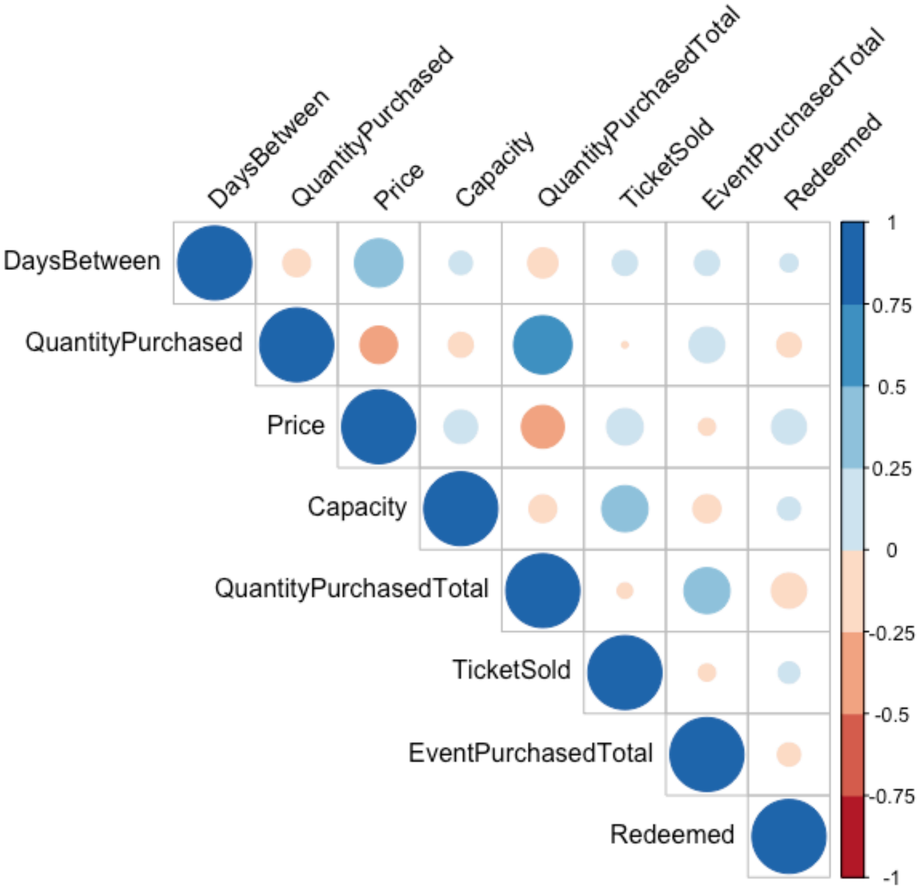

Figure 2.1 Correlation plot

Figure 2.1 gives the correlation plot between numeric variables. The largest value is 0.62 which is the correlation between QuantityPurchased and QuantityPurchasedTotal. All the other values are lower than 0.4. There is no big problem in correlation between numeric variables.

3. Data Visualization

3.1 Attribute Information

We consider 13 variables as our potential predictors in this case, 6 of which are categorical variables while the rest of which are numerical. All categorical variables are nominal. And among all numerical, all of them are discrete variables except “Capacity”. Here, we still will treat them as continuous ones.

We also created a Shiny Application (URL: https://wenkehuang.shinyapps.io/appss/) to help us visualize and present our visualization plots more efficiently. How to use this Shiny Application will be discussed briefly in 3.4.2 part.

3.2 Categorical variables vs Redemption rate

For all bar plots with categorical variables on x-axis, all columns are sorted from the one with the highest redemption rate to the one with the lowest. The percentages of being redeemed and not being redeemed are also labeled in the plots.

3.2.1 PriceType

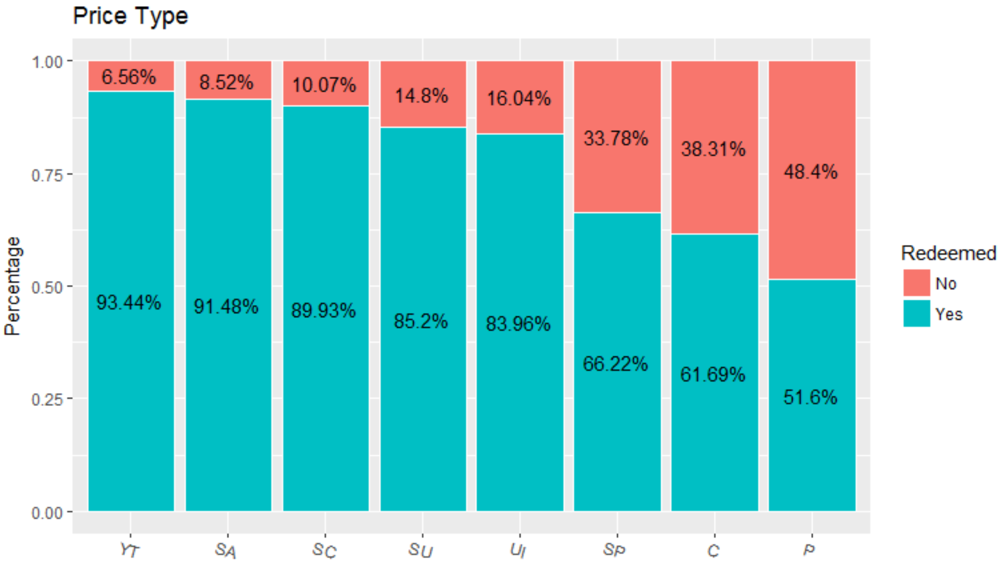

Figure 3.1 PriceType

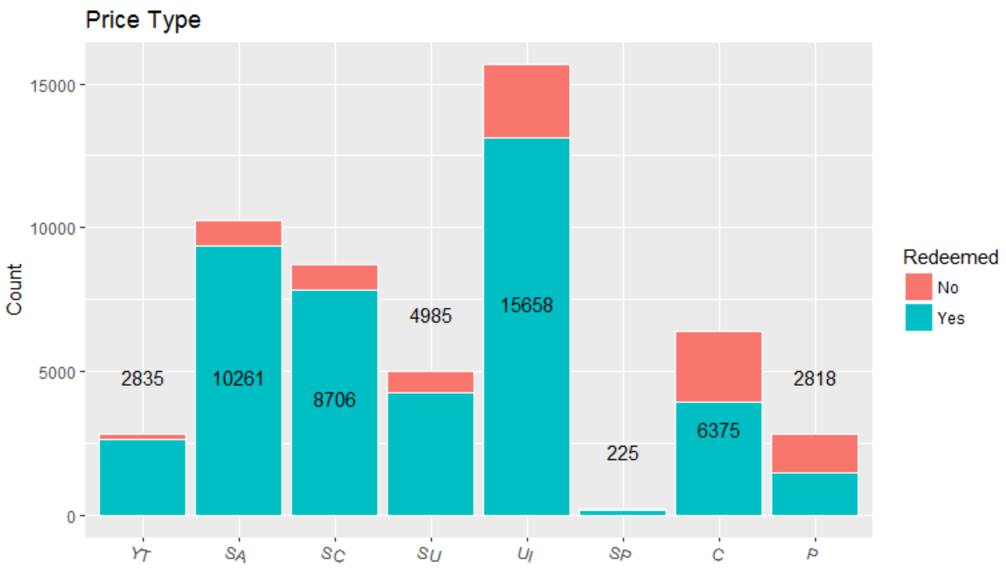

Figure 3.2 PriceType

It is obvious that the redemption rate of SP, C and P these three types are all lower than 67% while redemption rates of the rest five ticket types are all greater than 83%. Also, we can read from the plot that UI, SA and SC these three types take most of the tickets that were sold during the last season.

Since SP, C and P ticket take 0.43%, 12.29% and 5.43% of the total sold tickets, it would be more efficient if we focus on searching solutions of increasing the redemption rates of C and P tickets.

3.2.2 Theatre

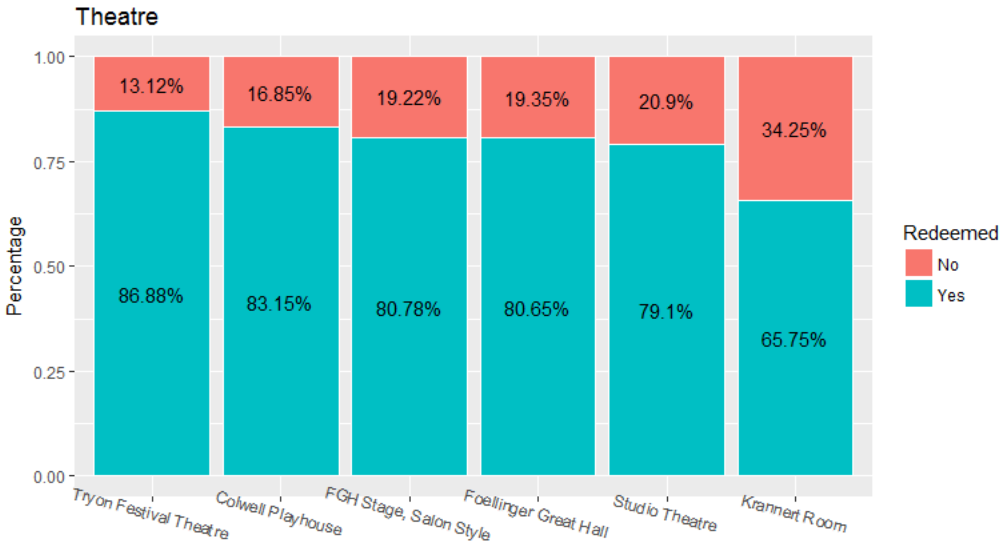

Figure 3.3 Theatre

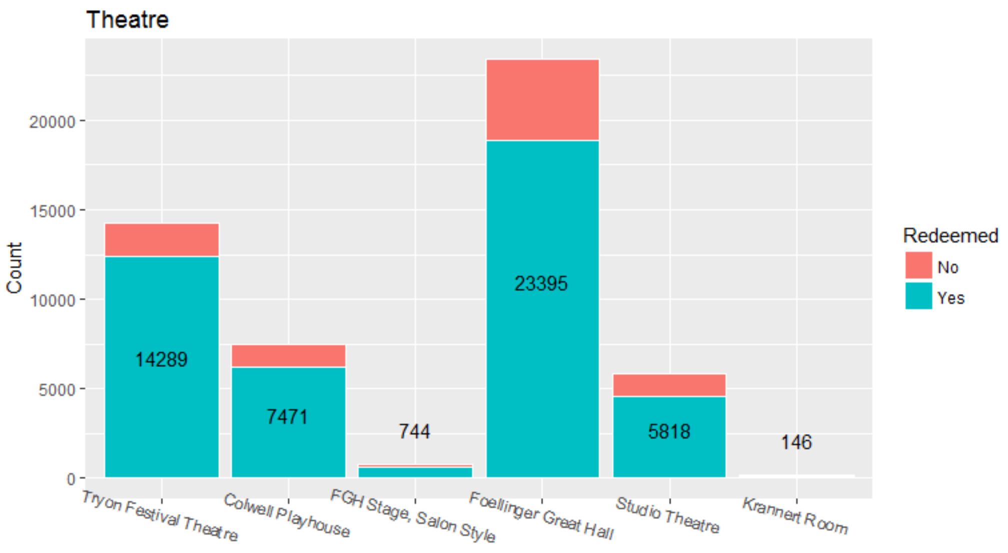

Figure 3.4 Theatre

We can notice that the redemption rate of Krannert Room shows only are the lower than 66% while the redemption rates of the rest five are all greater than 79%. At the same time, only 146 Krannert Room show tickets and 744 FGH Stage tickets were sold last year, and approximately half of the tickets were Foellinger Great Hall show tickets.

Hence, mainly focusing on increasing the redemption rates of Foellinger Great Hall and Studio Theatre is suggested here because of their popularity and potential ability, respectively.

3.2.3 Disposition

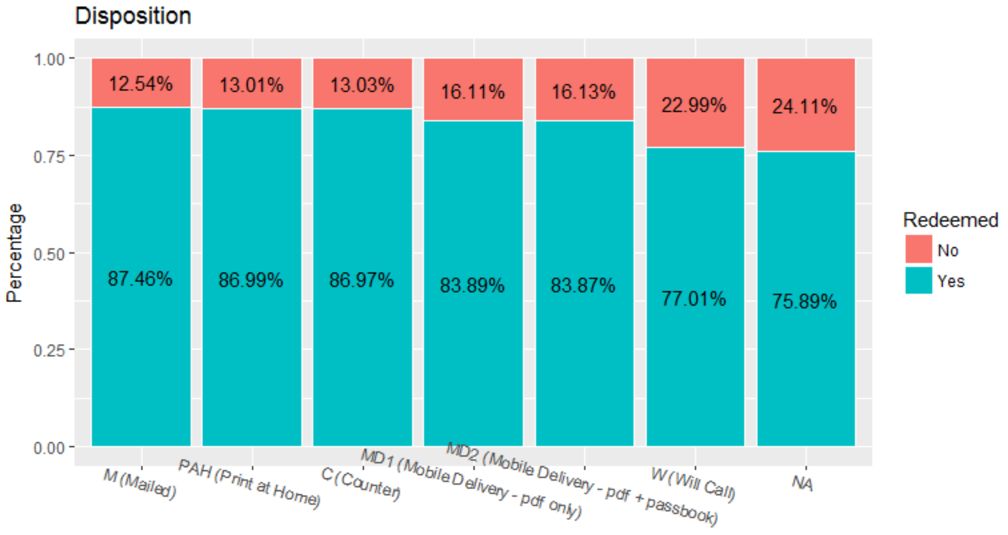

Figure 3.5 Disposition

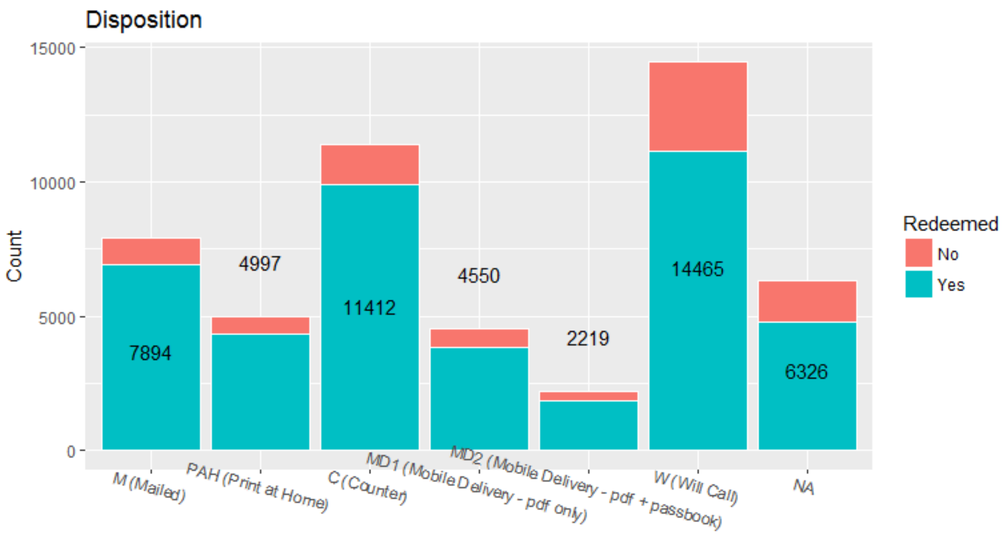

Figure 3.6 Disposition

The redemption rate of Will Call is the lowest among all none-missing records. The redemption rates of the rest five valid disposition methods are all around 86%. We can also read that most tickets were still sold by Will Call or purchased at the Counter.

Figuring out solutions of increasing the redemption rates of tickets purchased via Will Call is an efficient way to improve the overall redemption rates, because of its share of total tickets and current low redemption rate. Also, we notice that here missing values which were recorded as "NA" can be found here, and they are treated as a single category. If we could be able to find solutions to correct those records, our prediction can and will be more precise and reliable.

3.2.4 PurchasedOnline

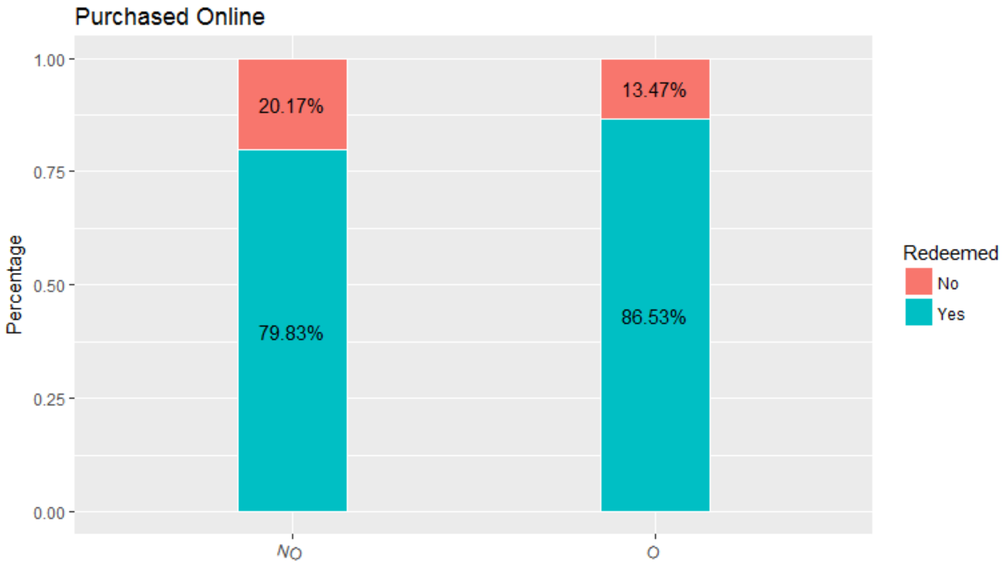

Figure 3.7 PurchasedOnline

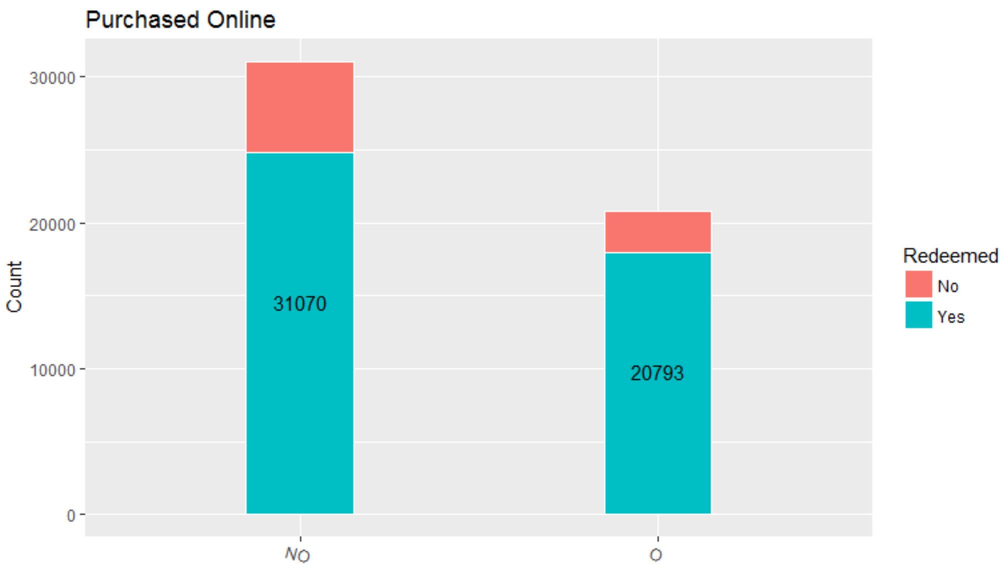

Figure 3.8 PurchasedOnline

We find that the Online Purchased tickets tend to have a higher redemption rate while the redemption rate of Offline Purchased tickets is a little bit lower. Most of tickets were still purchased via old ways.

It is obvious that trying to increase the redemption rate of Offline Purchased tickets should be what we focus on mainly. Because Offline Purchased tickets still take 59.91% of total tickets while the redemption rate is slightly lower than 80%.

3.2.5 Producer

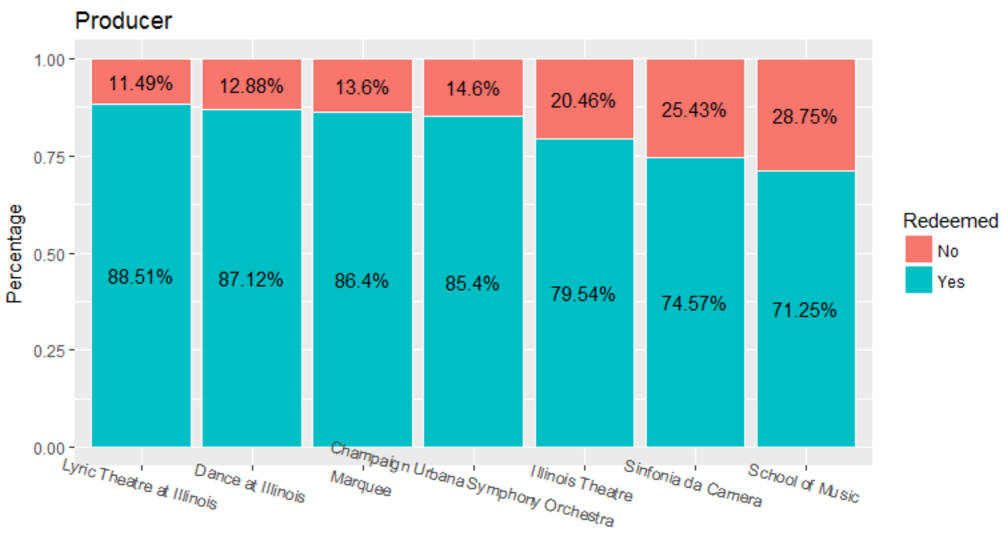

Figure 3.9 Producer

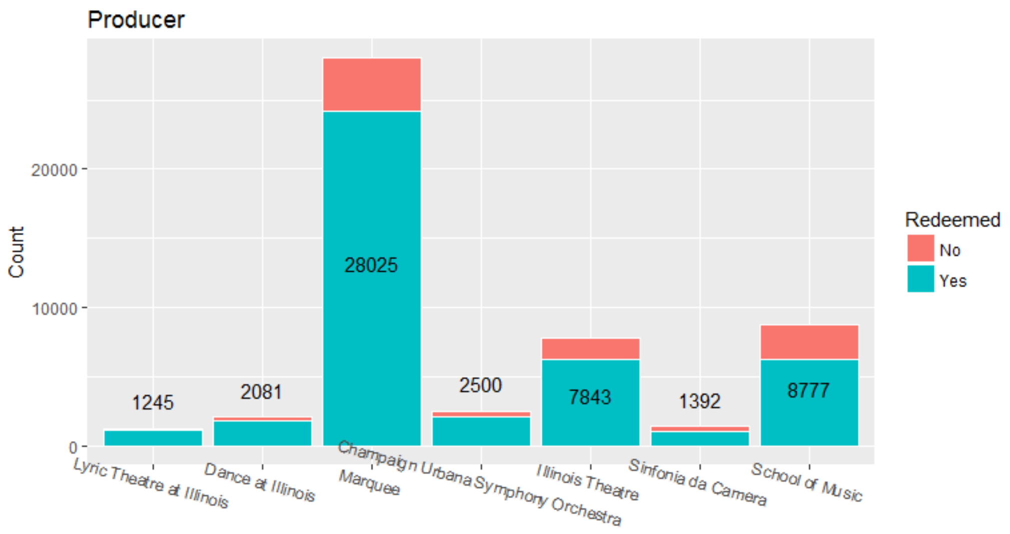

Figure 3.10 Producer

More than half (54.06%) of the tickets were sold are from producer Marquee. The redemption rates of Illinois Theatre, School of Music and Sinfonia da Camera are obviously lower than the redemption rates of the other producers.

We can calculate that Illinois Theatre, School of Music and Sinfonia da Camera take 15.12%, 16.92% and 2.68%, respectively. Theoretically, concentrating on the first two producers and trying to increase the redemption rates of these two should be an efficient way. However, the switching producers or making shows from one producer more attractive does not seem to be under the control of Krannert Center, which makes it harder than other suggestions.

3.2.6 Category

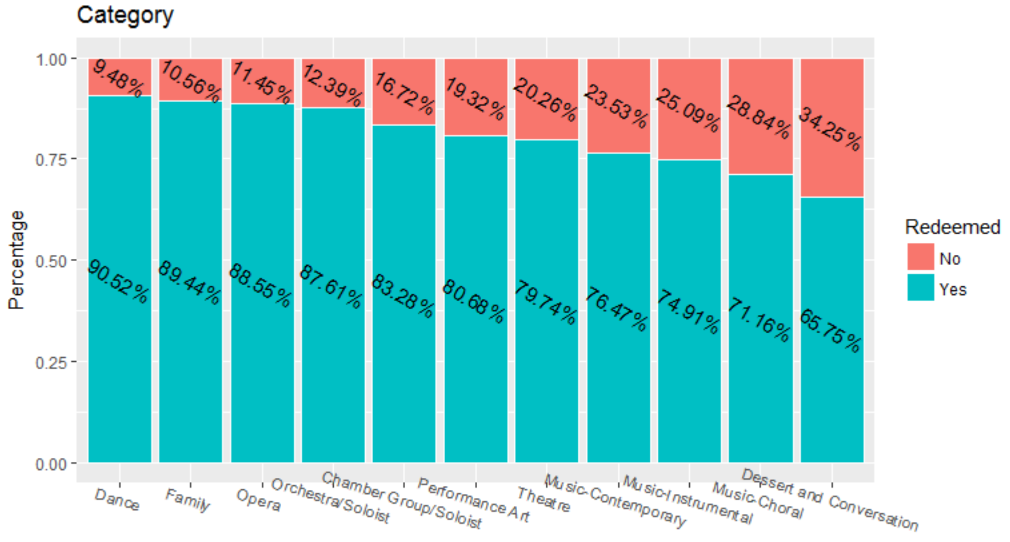

Figure 3.11 Category

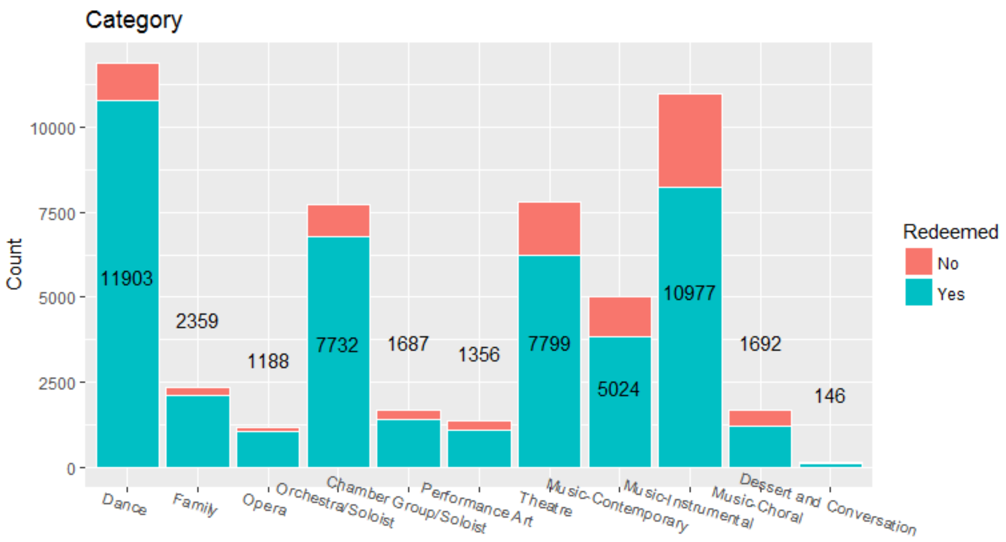

Figure 3.12 Category

Dance, Family, Opera and Orchestra/Soloist these 4 categories tend to have higher redemption rates. Dessert and Conversation, music related categories may be less attractive.

We can also notice that very few (0.28%) tickets are from Dessert and Conversation category, which means that focusing on music related categories can be more efficient.

3.3 Continuous variables vs Redemption rate

For all numerical variables, we will apply several cutoffs make it into categorical variables, and then draw percentage bar plot and histogram just like the previous part.
For the plots with user-defined categorical variables on x-axis, all columns are sorted from the lowest x value to the highest since they are ordinal categorical variables which is not like the previous part. The percentages of being redeemed and not being redeemed are also labeled in the plots.

3.3.1 TicketSold

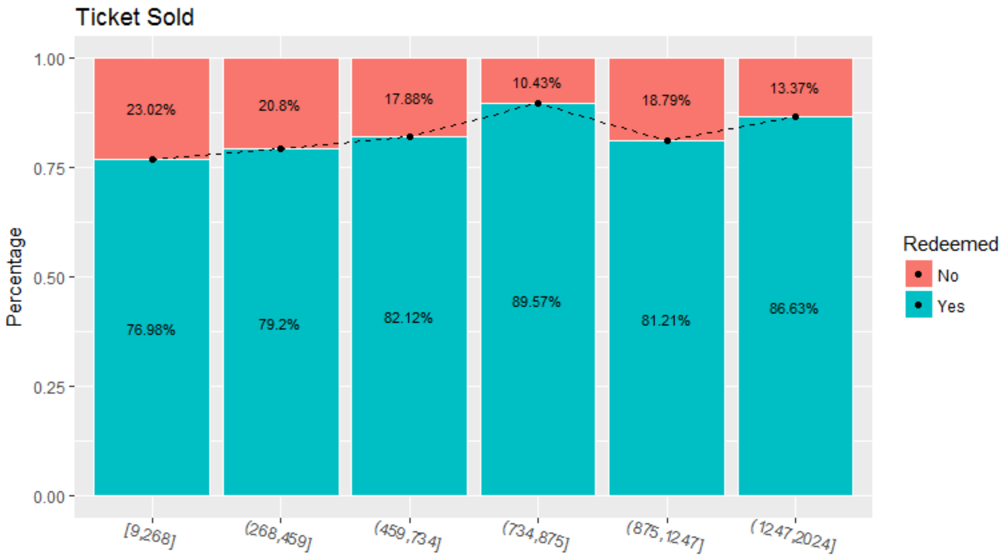

Figure 3.13 TicketSold

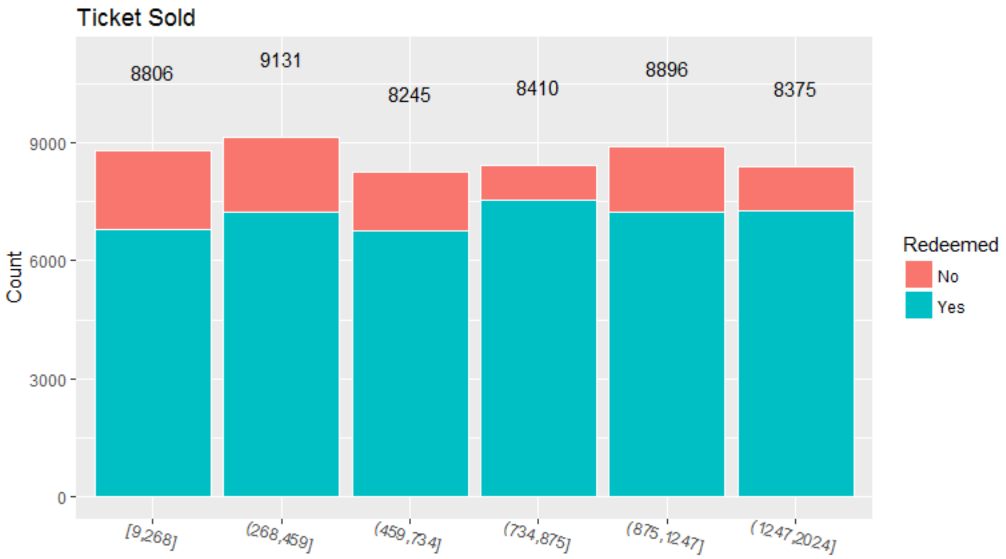

Figure 3.14 TicketSold

The range of variable "Ticket Sold" is [9, 2024]. All shows with Ticket Sold in (1247, 2024) must take place at Foellinger Great Hall because only this theatre has a great number of seats. We can read that the redemption rate will go up as Ticket Sold going up at the beginning. After Ticket Sold reaches roughly 900, the redemption rate will drop down a little.

It would be a good idea if we focus on the shows with lower and higher Tickets Sold and try to find out what affects their corresponding redemption rates.

3.3.2 Capacity

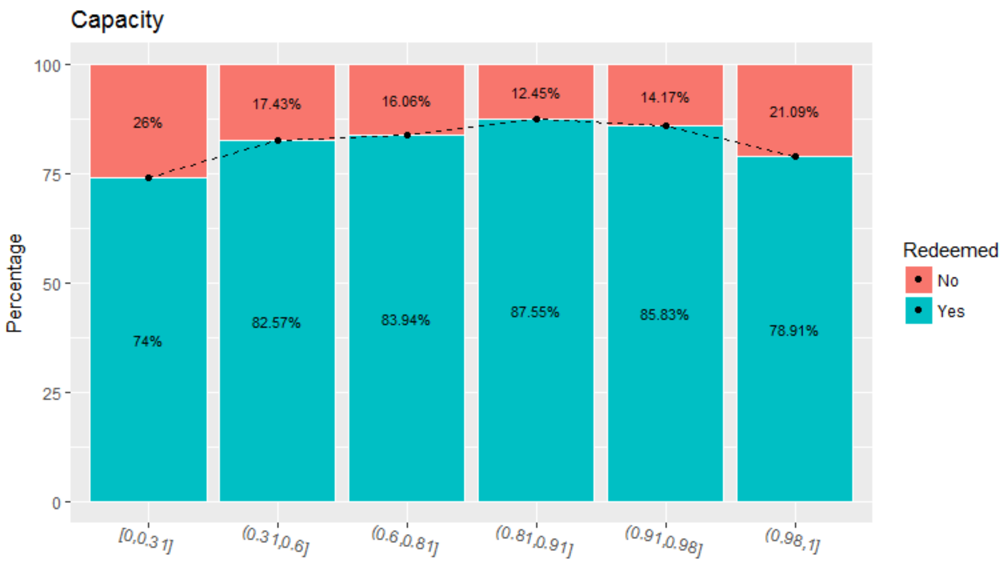

Figure 3.15 Capacity

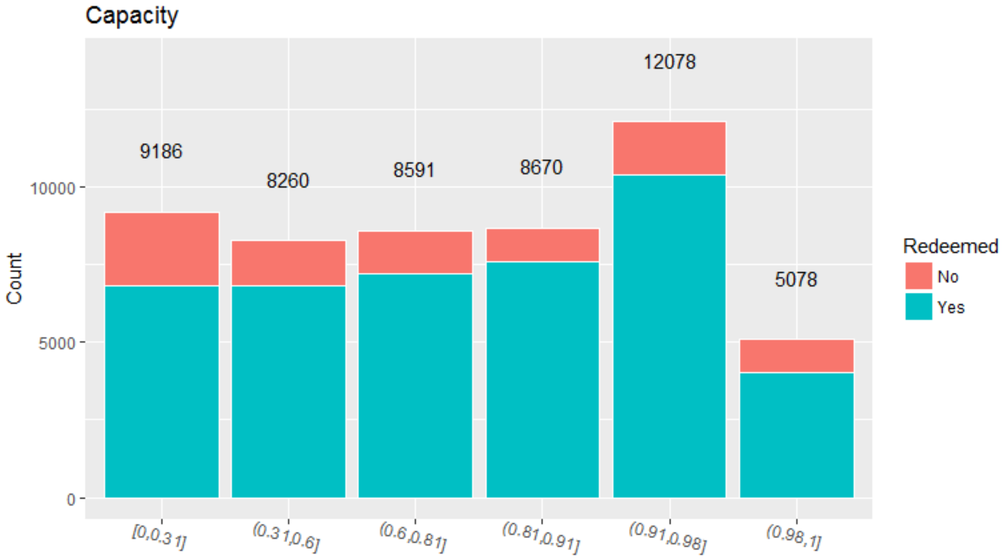

Figure 3.16 Capacity

The capacity is the only non-discrete continuous variable in our whole dataset. We can easily see the trend that the lower redemption rates locate at two sides while higher redemption rates are in the middle.

Hence, we should focus on the lower and higher Capacity shows mainly to increase our overall redemption rates. Also, because of this obvious trend mentioned in the previous paragraph, quadratic terms will be considered in the model fitting part.

3.3.3 Price

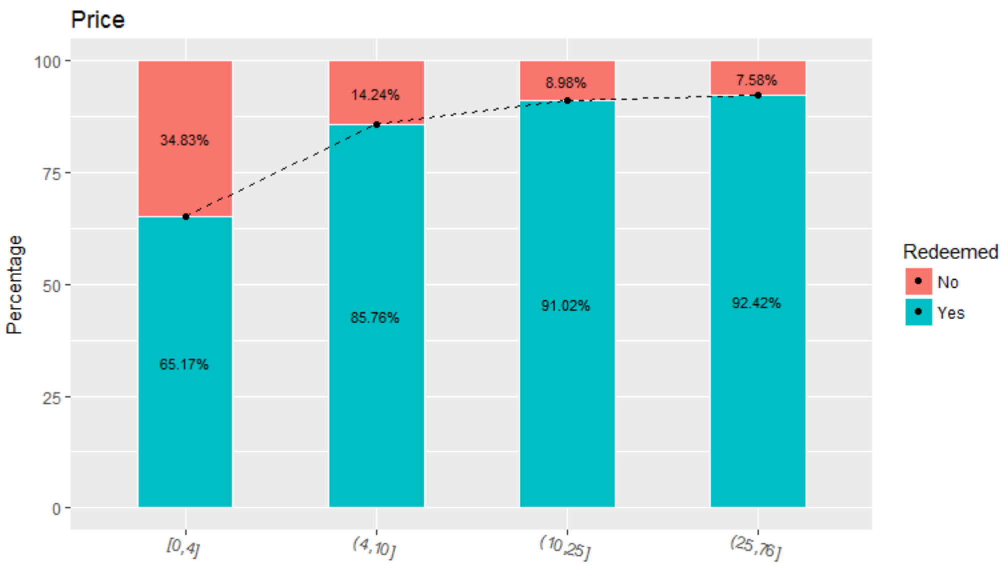

Figure 3.17 Price

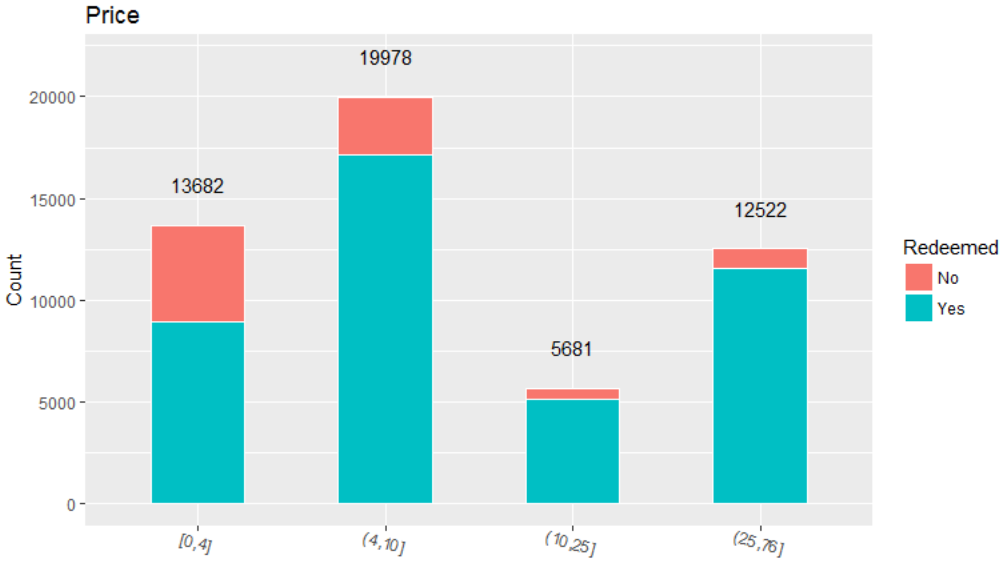

Figure 3.18 Price

It is obvious that the redemption rates will go higher as the ticket price going up, which is an easily understandable trend.

Because if the ticket price is very low or Krannert Center just sends out free tickets, customers may not care about them too much. On the other hand, if a customer is willing to spend a great amount of money on the show, for example 76 dollars, it means that the customer expects that one and no matter what happens he or she will come.

For the information that we have, Krannert Center sends out free tickets or offers low price tickets very often, and they take roughly 26.38% of the total. Raising the ticket price from [0, 4] to (4, 10) seems to be one plausible solution which may help increase the redemption rate.

3.3.4 DaysBetween

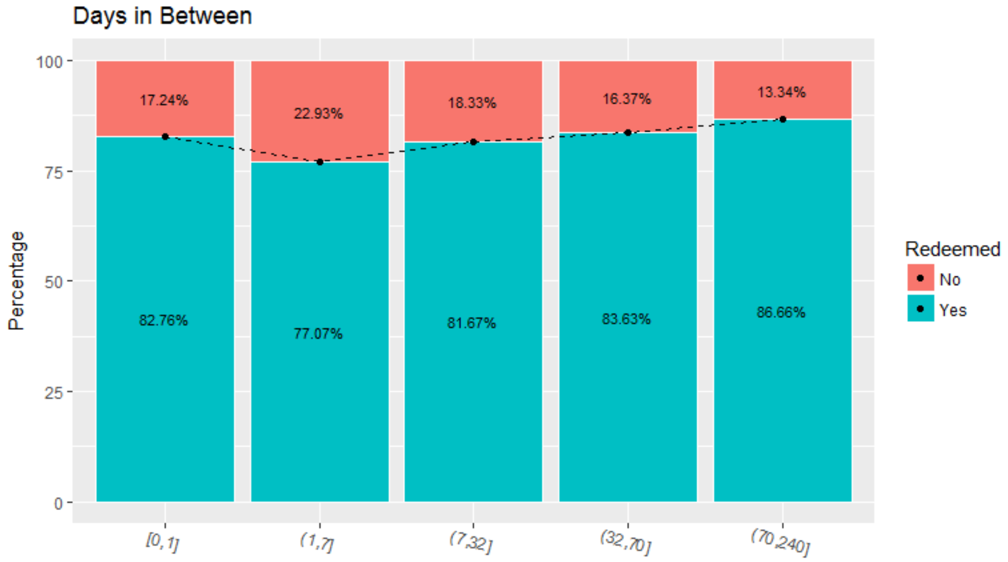

Figure 3.19 DaysBetween

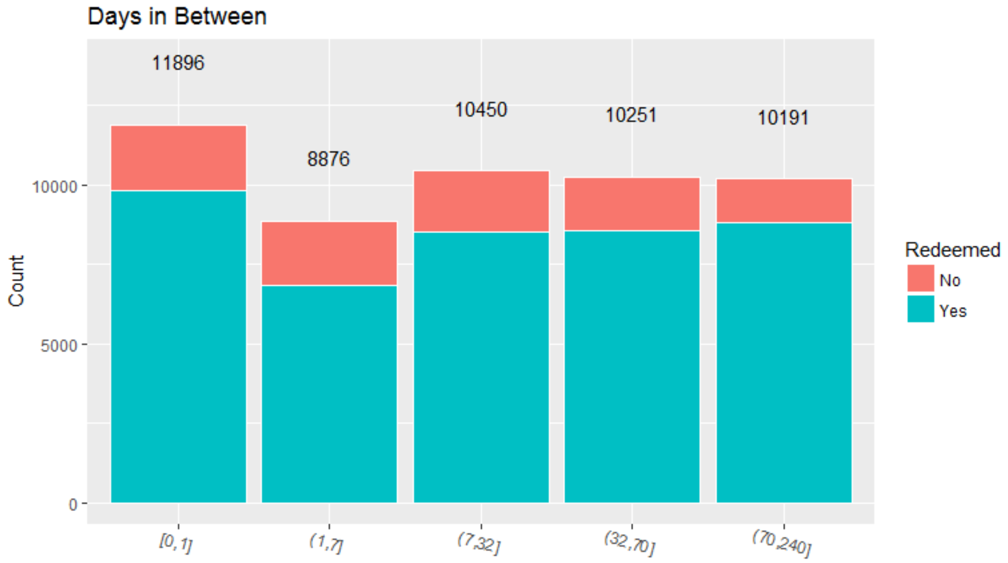

Figure 3.20 DaysBetween

We can read from the percentage bar plot that it has higher redemption rates at two sides and lower redemption rates in the middle.

This trend gives us some information that is against common sense. Because we read that the category (1, 7) has the lowest redemption rate, and as days in between going higher, the redemption rate goes up. Like Capacity above, quadratic terms will also be considered in the following model fitting part because of the trend the redemption rates.

3.3.5 QuantityPurchased

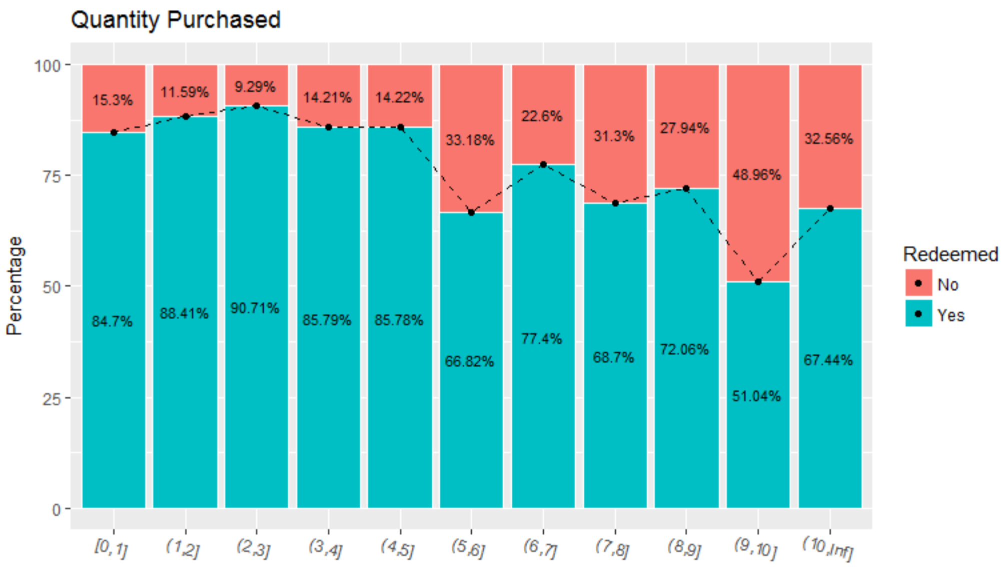

Figure 3.21 QuantityPurchased

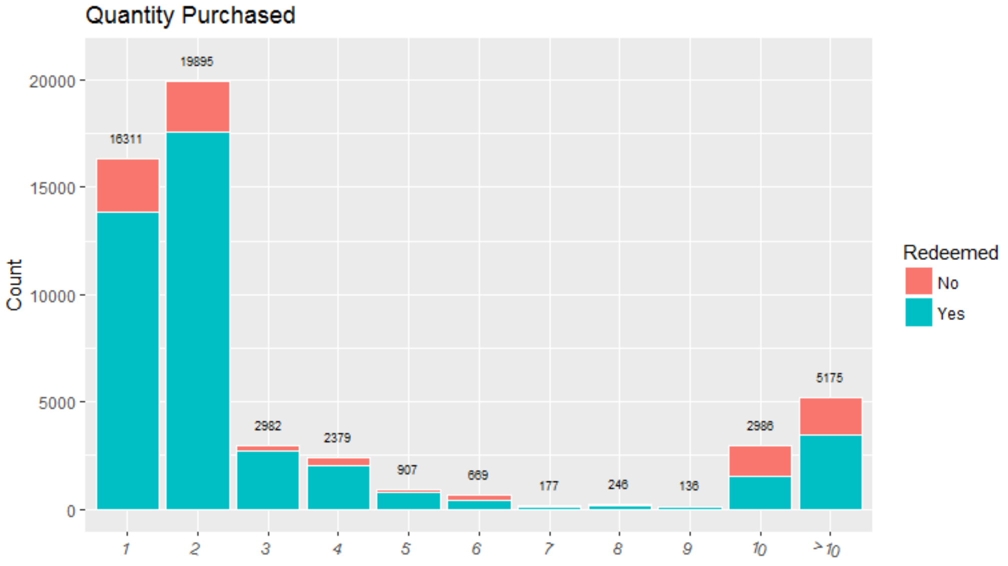

Figure 3.22 QuantityPurchased

Quantity Purchased stands for how many tickets each customer purchased for one event. We can read that most customers chose to purchase one or two tickets at one time, and at the same time we still can find some customers buying ten or more than ten tickets, too.

We can see that it has higher redemption rates in the left side. Redemption rate will also go down rapidly after Quantity Purchased reaches 6.

We read that customers who tend to buy many tickets at one time tend not to show up in the show, which leads to a lower redemption rates. Krannert Center can consider limiting the amount of tickets a customer can buy at one time to try to solve this issue.

3.3.6 QuantityPurchasedTotal

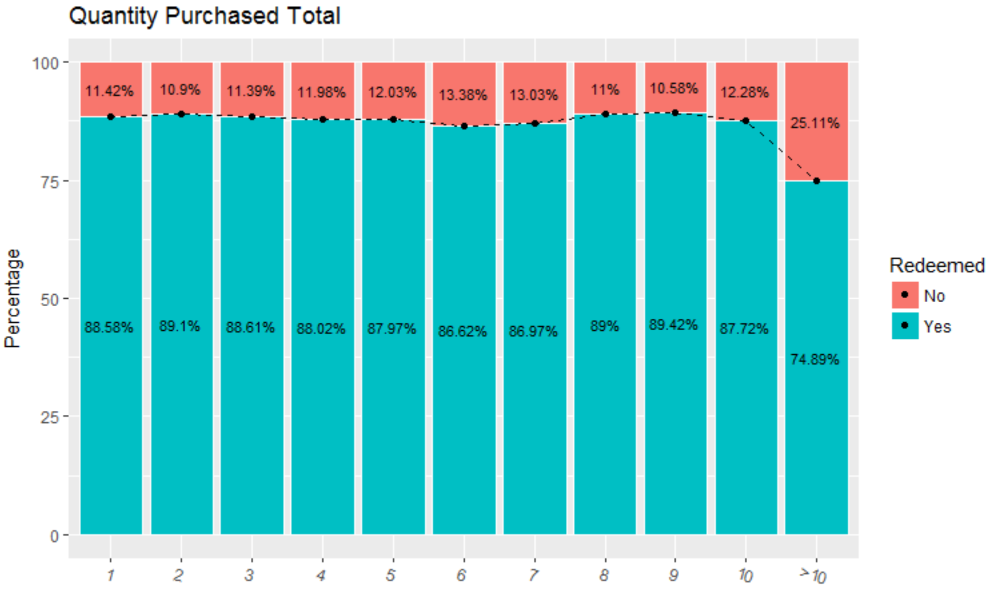

Figure 3.23 QuantityPurchasedTotal

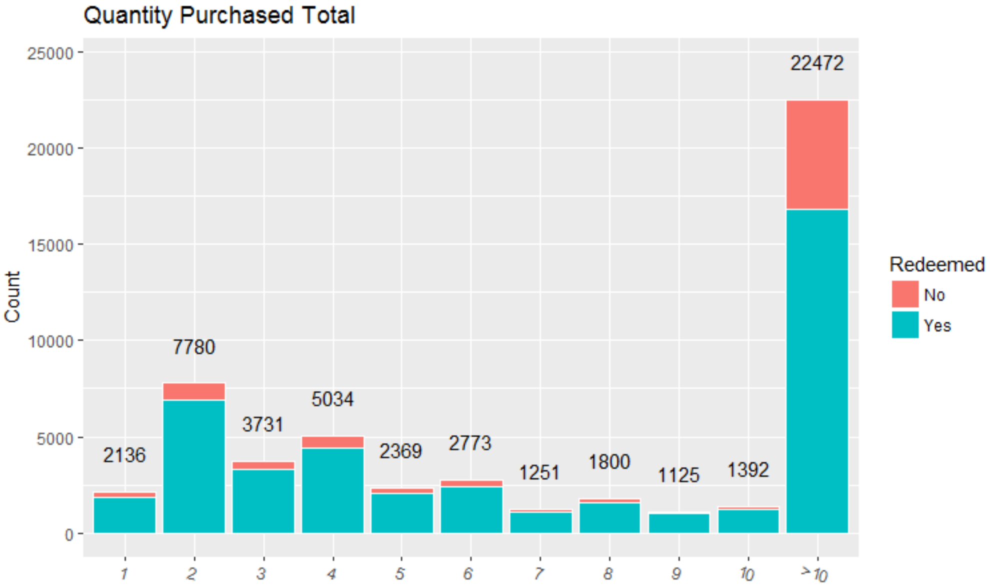

Figure 3.24 QuantityPurchasedTotal

Quantity Purchased Total summarizes how many tickets each customer purchased for the whole time. We can read that redemption rate goes down rapidly as Quantity Purchased Total passing 10, which means that frequent buyers tend to have lower redemption rates. We can still notice that most customers can be considered as frequent buyers.

Frequent buyers should be the group of customers that Krannert Center focus on mostly to help increase the overall redemption rate.

3.3.7 EventPurchasedTotal

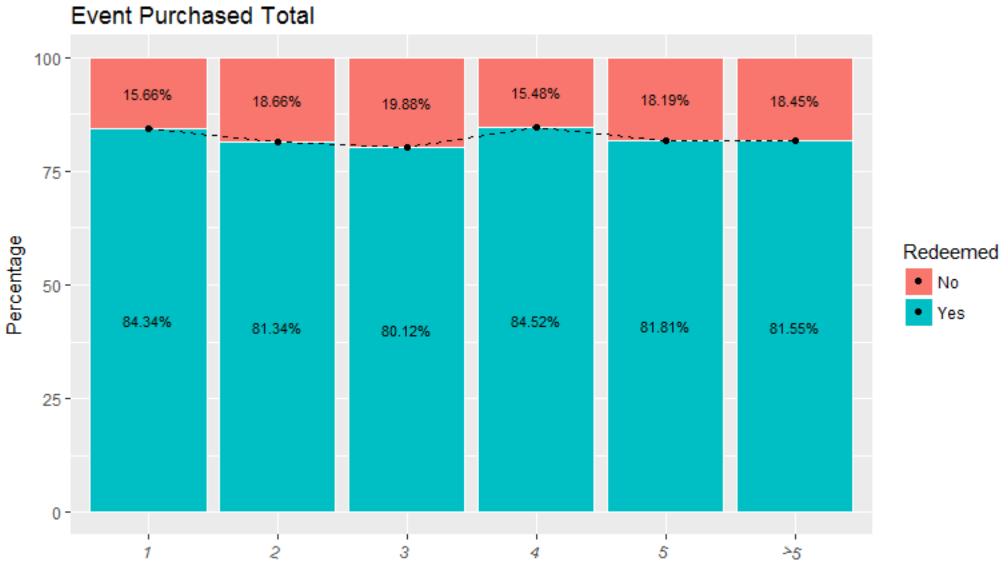

Figure 3.25 EventPurchasedTotal

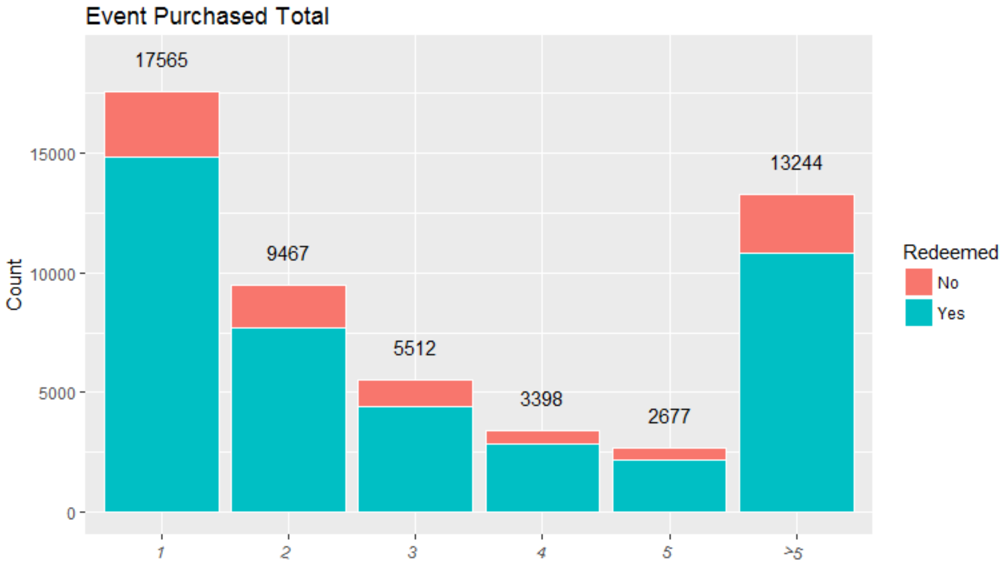

Figure 3.26 EventPurchasedTotal

Event Purchased Total stands for how many events did customers purchase during the whole time. Most customers chose to purchase one or two different events. We can also read that all redemption rates are around 83% in the percentage bar plot.

3.4 Predictors variables vs Predictor variables

3.4.1 Introduction

Apart from plots of predictor variable vs response variable, we also drew plots of predictor variable vs predictor variable to help find our whether any two predictor variable variables are correlated, which may generate multicollinearity problem in our following model fitting part.

- For categorical variable vs categorical variable, percentage bar plot is drawn just like the previous two parts, and Pearson Chi-Square test is applied to quantitively test if the result is significant or not.
- For numerical variable vs numerical variable, scatter point plot is drawn, and Pearson Correlation Coefficient is calculated to help quantitively test the correlationship between them.
- For numerical variable vs categorical variable, box plot is drawn, and ANOVA table and its corresponding Tukey HSD test is applied here to help us test quantitatively.

3.4.2 Shiny Visualization Application

Here, we have 13 predictor variables in total. To present the plots that we drew more efficiently, we chose to create a Shiny Application (URL: https://wenkehuang.shinyapps.io/appss/) to help show and present our results. By clicking and choosing different "Variable 1" and "Variable 2", any visualization plots will be shown immediately. All percentage bar plots above can also be found from our Shiny Application.

If "Variable 1" and "Variable 2" stand for the same variable, the histogram of the corresponding variable will be shown and not statistical test will be applied.

4. Statistical Tests

4.1 Introduction and method

In this section, we use Pearson's chi-square test and Tukey's test to do some statistical inference on the data and analyze the relationship between the predictors and response variable Redeemed. The variable we analyzed in this part include variables related to customer: PriceType, Disposition, PurchasedOnline and variables related to event: Theatre, Producer and Category.

First, we want to know if there is significant correlation between the variable we test and the response variable redeemed. We used Pearson's chi-square test to testify our assumption. For each variable, the null hypothesis would be: there is no significant relationship between that variable and the response variable. The p-value for all 6 variables we test are very small but in this case, the p-value cannot provide much information. Since we have a relatively large sample size, even tiny correlation tends to be significant in the test.

Next, we use Tukey's test to see the pairwise comparison within each variable. The formula for Tukey's test is:

$$q_s = \frac{Y_A - Y_B}{SE}$$

Where Y_A and Y_B are the two means been compared, SE is the standard error of the data. Since we build a generalized linear model to do the test and the p-values from chi-square test are all less than 0.05, so the assumptions of Tukey's test can be satisfied.

In the following subsections, we present the result of Tukey's test. In each table, the group column indicates the relationship between different levels in each variable. If two levels are from the same group, that means the mean value of redemption rate among those two levels are not significantly different.

4.2 PriceType

Table 4.1 Tukey's test on PriceType

	PriceType	**Mean(Redeemed)**	**Group**
1	YT	0.934	a
2	**SA**	**0.915**	**ab**
3	**SC**	**0.899**	**b**
4	SU	0.852	c
5	**UI**	**0.839**	**c**
6	SP	0.662	d
7	**C**	**0.617**	**d**
8	P	0.516	e

SA, SC, UI and C are four major price type with relatively higher percentage. SA stands for standard admission and SC stands for senior citizen, these are regular price tickets. UI stands for tickets for UI student and faculty, these are discount tickets. Price type C are free tickets.

In those 4 major levels, price type C has the lowest redemption rate, that make sense because it's free tickets. And redemption rate for UI tickets are relatively lower than regular price tickets.

4.3 Disposition

Table 4.2 Tukey's test on Disposition

	Disposition	Mean(Redeemed)	Group
1	M (Mailed)	0.875	a
2	PAH (Print at Home)	0.870	a
3	**C (Counter)**	**0.870**	**a**
4	MD1 (Mobile Delivery - pdf only)	0.839	b
5	MD2 (Mobile Delivery - pdf + passbook)	0.839	b
6	**W (Will Call)**	**0.770**	**c**

From the table, we can see that there is no big difference between first 5 levels, the confidence interval between pairs from group a and b are close to 0. Krannert center now provide print at home and mobile tickets for all kinds of events, this may help increase the redemption rate since there is significantly difference between level will call and print home or mobile tickets.

4.4 PurchasedOnline

Table 4.3 Tukey's test on PurchasedOnline

	PurchasedOnline	Mean(Redeemed)	Group
1	O	0.865	a
2	NO	0.798	b

From the result of Tukey's test, there is significant difference between the two-different purchase method, the redemption rate for tickets purchased online are slightly higher.

4.5 Theatre

Table 4.4 Tukey's test on Theatre

	Theatre	Mean(Redeemed)	Group
1	**Tryon Festival Theatre**	**0.869**	**a**

2	**Colwell Playhouse**	**0.831**	**b**
3	FGH Stage, Salon Style	0.808	bc
4	**Foellinger Great Hall**	**0.807**	**c**
5	Studio Theatre	0.791	c
6	Krannert Room	0.658	d

The biggest theatre Foellinger has a relatively lower redemption rate compared to other two major theatre Tryon and Colwell Playhouse. Theatre Foellinger has more than 2000 seats. One possible reason is that besides some popular event like lang-lang or Chicago orchestra symphony, there are some not so popular event in this big theatre, this may lower the redemption rate in Foellinger.

4.6 Producer

Table 4.5 Tukey's test on Producer

	Producer	Mean(Redeemed)	Group
1	Lyric Theatre at Illinois	0.885	a
2	Dance at Illinois	0.871	a
3	**Marquee**	**0.864**	**a**
4	C-U Symphony Orchestra	0.854	a
5	**Illinois Theatre**	**0.795**	**b**
6	Sinfonia da Camera	0.746	c
7	**School of Music**	**0.713**	**d**

Producer can be used to group all kinds of events. From the result, we can see that the event produced by School of music has relatively lower redemption rate compared with other two major producers.

4.7 Category

Table 4.6 Tukey's test on Category

	Category	Mean(Redeemed)	Group
1	**Dance**	**0.905**	**a**
2	Family	0.894	ab
3	Opera	0.886	ab
4	**Orchestra/Soloist**	**0.876**	**b**
5	Chamber Group/Soloist	0.833	c
6	Performance Art	0.807	cd
7	**Theatre**	**0.797**	**d**

8	Music-Contemporary	0.765	e
9	**Music-Instrumental**	**0.749**	**e**
10	Music-Choral	0.712	f
11	Dessert and Conversation	0.658	f

Category is another variable we can use to group events. There are four main type of event: dance, orchestra, theatre and music.

We can see that all 3 music related categories have significantly lower redemption rate than other type of event and about 21 percent of the events are from music-instrumental category.

5. Prediction models

We include 13 predictors to construct prediction models:

- Capacity
- Category
- DaysBetween
- Disposition
- EventPurchasedTotal
- Price
- PriceType
- Producer
- PurchasedOnline
- QuantityPurchased
- QuantityPurchasedTotal
- Theatre
- TicketSold

As for methods, we perform 1) logistic regression and 2) random forest to predict the probability of redemption.

Throughout our analysis, we used 80% of the observations as the training dataset to train our models and use the rest 20% of the observations as the test dataset to evaluate the model fitting.

To measure the predictive power of our models, we use criteria of classification table and ROC curve. With classification calculated, we also report the sensitivity and specificity of our models.

5.1 Logistic regression

We first use group lasso to perform variable selection. Using the selected variables, we propose three logistic regression models, an ordinary one (Model 1), one with interaction terms of Theatre (Model 2) and one taking PriceType as an offset (Model 3).

5.1.1 Variable selection

There are many categorical predictors in our dataset like PriceType, Theatre, Disposition, PurchasedOnline, Producer, and Category. Those categorical predictors are coded as a collection of binary covariates in prediction models. When performing variable selection, we want to ensure that all the variables encoding the same categorical covariate are either included or excluded from the model together. For this purpose, we choose to use group lasso to do variable selection which can guarantee that our requirement can be satisfied.

After variable selection by group lasso, the following 9 predictors out of 13 are kept in our model:

- DaysBetween
- EventPurchasedTotal
- Price
- PriceType
- Producer
- QuantityPurchased
- QuantityPurchasedTotal
- Theatre
- TicketSold

Notice that the three new predictors we create, QuantityPurchased, QuantityPurchasedTotal, and TicketSold, are all kept.

And the following 4 predictors out of 13 are eliminated from our model:

- Capacity
- Category
- Disposition
- PurchasedOnline

So, the 4 predictors Capacity, Category, Disposition, and PurchasedOnline are not so powerful in predicting probability of redemption with the other 9 predictors present in logistic regression model.

5.1.2 Logistic regression Model 1

We use the selected 9 predictors to construct a logistic regression model. We also try to include the quadratic terms of DaysBetween and QuantityPurchased as suggested by data exploration. It turns out that quadratic term of DaysBetween is statistically significant while that of QuantityPurchased is not.

Table 5.4 Parameter estimates of logistic regression Model 1

	Estimate	P-value		Estimate	P-value
Intercept	-6.66e-01	1.00e-11	P Dance at Illinois	4.98e-01	3.89e-05
PriceTypeP	-8.25e-02	2.64e-01	P Illinois Theatre	-2.60e-01	9.80e-03
PriceTypeSA	1.60e+00	2.11e-115	P Lyric Theatre	5.69e-01	3.40e-05
PriceTypeSC	1.44e+00	5.92e-77	P Marquee	9.56e-02	2.17e-01
PriceTypeSP	-3.73e-02	8.18e-01	P School of Music	-2.30e-01	8.58e-03
PriceTypeSU	1.29e+00	1.66e-88	P Sinfonia da Camera	-2.15e-01	4.26e-02
PriceTypeUI	1.14e+00	1.73e-137	DaysBetween	-1.03e-02	6.22e-31
PriceTypeYT	1.75e+00	8.02e-76	QuantityPurchased	6.20e-03	9.57e-09
T FGH Stage	-2.13e-01	8.68e-02	Price	1.50e-02	4.27e-14
T FGH	2.22e-02	7.49e-01	QuantityPurchasedTotal	-1.18e-03	8.42e-52

T Krannert Room	-1.54e+00	1.82e-13	TicketSold	1.43e-04	1.20e-03
T Studio Theatre	7.11e-02	2.27e-01	EventPurchasedTotal	-1.42e-02	9.95e-17
T Tryon Festival	1.38e-01	3.84e-02	DaysBetween^2	3.78e-05	5.52e-14

Every predictor in Model 1 is statistically significant. Table 5.1 gives the parameter estimates of Model 1. To interpret the estimates, for example for Price, the estimate is 0.0015, which means the odds of redemption multiply by exp(0.0015) = 1.0015 for every one-unit increase in Price. In general, the probability of redemption will go up along with the increase in the predictors with positive estimates and go down along with the increase in the predictors with negative estimates. So, the predictors with positive estimates have a positive relationship with the probability of redemption and the predictors with negative estimates have a negative relationship with the probability of redemption.

Table 5.5 Classification table of logistic regression Model 1

	Prediction, $\pi_0 = 0.826$		Prediction, $\pi_0 = 0.500$		
Actual	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	Total
Redeemed	6383	2112	8202	293	8495
Unredeemed	692	1146	1497	341	1838

We use two different cutoff value π_0, 0.826 (mean redemption rate of the training data) and 0.5, to construct the classification table. With the classification table, we can calculate the sensitivity and specificity. Sensitivity measures the proportion of unredeemed tickets that are correctly identified (i.e. the percentage of unredeemed tickets that are correctly identified as unredeemed) and specificity measures the proportion of redeemed tickets that are correctly identified.

When $\pi_0 = 0.826$, we have

- sensitivity is 62%
- specificity is 75%

When $\pi_0 = 0.5$, we have

- sensitivity is 19%
- specificity is 97%

As we can see, when using more conservative cutoff value π_0, we can get better specificity but worse sensitivity.

To better summarize the predictive power for all possible π_0, we'd also like to use receiver operating characteristic (ROC) curve to measure the performance of our prediction model. The criterion of ROC curve is that the greater the area under the ROC curve, the better the predictive power of the model.

Figure 5.1 gives the ROC curve of logistic regression Model 1, and the area under the curve is 0.7433, which is fair though not great.

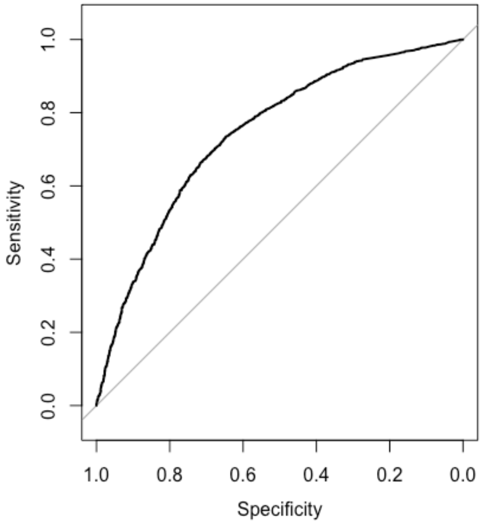

Figure 5.1 ROC curve of logistic regression Model 1

5.1.3 Logistic regression Model 2

In the second logistic regression model, we include the interaction term of Theatre with other predictors. So, compared with logistic regression Model 1, this is a more complicated model in the sense that it has much more parameters to estimate.

Table 5.6 Likelihood ratio test between Model 1 and Model 2

	Resid. Df	Resid. Dev	Df	Dev	P-value
Model 1	41305	33531			
Model 2	41238	32923	67	607.73	3.90e-88

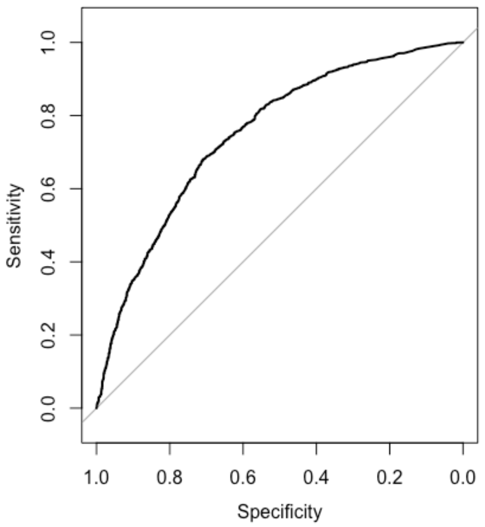

Figure 5.2 ROC curve of logistic regression Model 2

We perform likelihood ratio test to compare Model 1 and Model 2. From Table 3, the p-value is 3.90e-88. So statistically, we'd prefer Model 2 rather than Model 1.

Figure 5.2 gives the ROC curve of logistic regression Model 2, and the area under the curve is 0.75, which is pretty close to that of Model 1. We can see that, compared with Model 1, the predictive power of Model 2 does not increase too much. At the same time, Model 2 includes 67 more parameters than Model 1, which make it a much more complicated model. So for the purpose of simplicity, Model 1 is still preferred in terms of prediction.

However, the significance of Model 2 lies in that it is actually equivalent to fit separate model for each theatre. In that way, we can include more predictors in our model, especially some theatre-specific predictors like Seatblock. Theatre-specific predictors like Seatblock cannot be included in Model 1 because the coding of the Seatblock is different for different theatres.

Here, we don't include Seatblock in Model 2 because we find out that unfortunately, the coding of the Seatblock is different even for the same theatre. Model 2 may be useful in future analysis when Seatblock can be modified or more theatre-specific predictors can be recorded.

5.1.4 Logistic regression Model 3

In later analysis, we find out that DaysBetween is a key factor in predicting probability of redemption (discussed in section 3.2.2). From data exploration, we also find out that tickets of PriceType with lower redemption rate like P, C and SP usually have small value of DaysBetween. This may mislead to the conclusion that tickets with small value of DaysBetween would less likely to be redeemed, which does not make much sense.

We suspect PriceType to be a confounding variable and want to know the effect of the other predictors on predicting the probability of redemption conditioning on the effect of PriceType. So we take PriceType as an offset and construct our logistic regression model.

Table 5.7 Parameter estimates of logistic regression Model 3

	Estimate	P-value		Estimate	P-value
Intercept	1.63e-01	8.83e-02	P Dance at Illinois	5.37e-01	6.33e-06
T FGH Stage	-2.41e-01	5.40e-02	P Illinois Theatre	-2.24e-01	2.01e-02
T FGH	4.42e-02	5.26e-01	P Lyric Theatre	6.15e-01	5.29e-06
T Krannert Room	-1.77e+00	2.60e-18	P Marquee	1.16e-01	1.24e-01
T Studio Theatre	8.17e-02	1.64e-01	P School of Music	-1.07e-01	1.89e-01
T Tryon Festival	1.24e-01	5.97e-02	P Sinfonia da Camera	-2.08e-01	4.98e-02
DaysBetween	-1.04e-02	2.79e-33	QuantityPurchasedTotal	-1.01e-03	8.38e-43
QuantityPurchased	4.78e-03	3.51e-06	TicketSold	1.45e-04	1.07e-03
Price	8.22e-03	1.16e-09	EventPurchasedTotal	-1.75e-02	1.06e-26
DaysBetween^2	3.77e-05	1.43e-14			

Table 5.4 gives the parameter estimates of Model 3. Compare the parameter estimates of Model 1 and Model 3, we can see that though the values change, the sign of each parameter does not change, which means the direction of the relationship does not change.

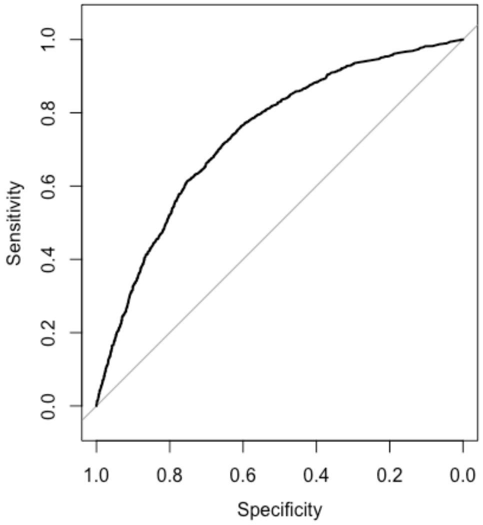

Figure 5.3 ROC curve of logistic regression Model 3

Figure 5.3 gives the ROC curve of logistic regression Model 3, and the area under the curve is 0.7389, which is still very close to previous ones.

5.2 Random Forest

We include all the 13 predictors to construct our random forest model.

Table 5.8 Classification table of random forest model

	Prediction, $\pi_0 = 0.835$		Prediction, $\pi_0 = 0.500$		
Actual	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	Total
Redeemed	5819	1756	7270	305	7575
Unredeemed	334	1190	855	669	1524

As before, we use two different cutoff value π_0, 0.835 (mean redemption rate of the training data) and 0.5, to construct the classification table.

When $\pi_0 = 0.835$, we have

- sensitivity is 78%
- specificity is 77%

When $\pi_0 = 0.5$, we have

- sensitivity is 44%
- specificity is 96%

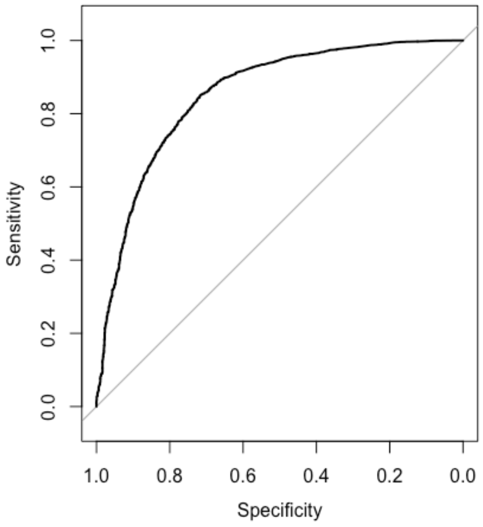

Figure 5.4 ROC curve of random forest model

Figure 5.4 gives the ROC curve of random forest model, and the area under the curve is 0.8529, which improves a lot from that of logistic model.

We can see that compared with logistic regression model, random forest model has a much stronger predictive power. And besides the strong predictive power, the advantage of random forest model also lies in that it provides a measure of the importance of predictors in predicting probability of redemption.

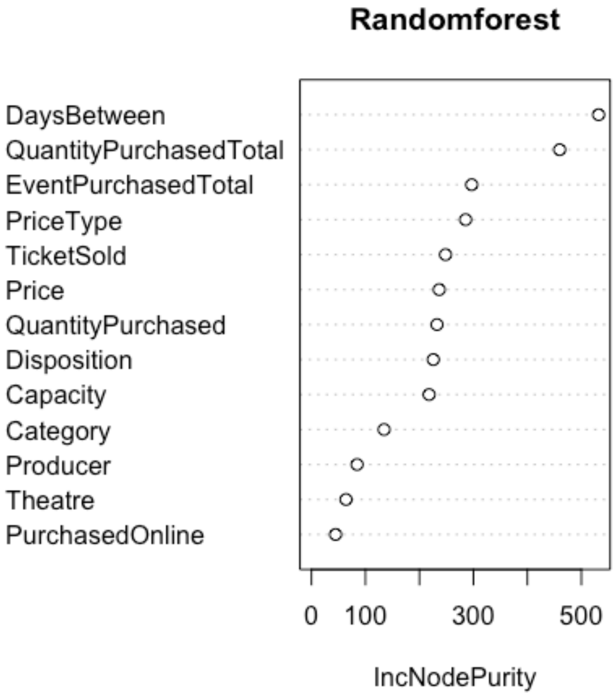

Figure 5.5 Importance plot of random forest model

In Figure 5.5, predictors are ordered according to the importance in predicting the probability of redemption. As we can see, DaysBetween is identified as the most important predictor, that is to say, DaysBetween is the most contributive predictor in predicting the probability of redemption in random forest model. However, the problem is that random forest model does not tell us anything about the relationship between the predictors and the response. For example, when the price of the ticket goes up, we don’t know how the probability of redemption would change accordingly from the random forest model. We can resort to our logistic regression model for help and fortunately, we can see that the key predictors selected by the random forest model like DaysBetween, QuantityPurchasedTotal, etc. are all kept in our logistic regression model by group lasso.

6. Discussion

6.1 Relationship between delivery option and redemption rates

From our analysis, there is statistically significant difference between different groups of delivery option. However, the predictor Disposition is eliminated from our logistic regression model by group lasso. Also, from the random forest model, Disposition is not identified as the key variable in predicting the probability of redemption.

So, we'd conclude that Disposition is not powerful in predicting the probability of redemption. Everyone would agree that providing more delivery options for customer is a good thing and a step forward. It may help improve the sales of tickets, but it does not help too much in predicting whether the ticket would be redeemed or not.

6.2 Relationship between purchase behaviors and redemption rates

- Purchased Online

  Tickets purchased online are more likely to be redeemed than those purchased offline.

- Frequent buyer

  Large values of QuantityPurchasedTotal or EventPurchasedTotal suggest that the customer is a frequent buyer. Compared to one-time buyers, frequent buyers are less likely to redeem their tickets.

- PriceType

  Tickets of PriceType YT, SA and SC are very likely to be redeemed.
  Tickets of PriceType SP, C and P are less likely to be redeemed.

6.3 Risky signal

From the random forest model, we identify three key factors in predicting whether the ticket would be redeemed or not. And from the parameter estimates of our logistic regression model, we can know what values of these key variables may indicate potential unredemption.

- DaysBetween

  Tickets bought just before the event day or a long time before the event day are more likely to be redeemed. More attention should be paid on the tickets bought in between.

- QuantityPurchasedTotal/EventPurchasedTotal

Frequent buyers indicated by large value of QuantityPurchasedTotal or EventPurchasedTotal are less likely to redeem their tickets. So if a ticket is purchased by someone that has bought a lot of tickets or has come to many events before, it is less likely to be redeemed.

- PriceType

  Tickets of PriceType SP, C and P are less likely to be redeemed.

7. Appendix

All R codes refer to the attachment ‘STAT427 Code_Krannert Center.txt’.