Capstone Project 2 Milestone Report

Subscription Box - A Trend or a Fad?

Mark Zhao

04/02/2020

Executive Summary

Leveraging proprietary data from my co-founded children's book subscription box business, I analyzed the key factors that predict the cancellations and also designed a model to predict the cancellation rate for existing active customers with 98% F1 score and 0.98 AUC. The key takeaways include:

- People are more likely to cancel the box around the renewal dates and more likely to subscribe before the monthly cut off days.

- Customers are smart and sensitive to the quality and quantity of the books in each month's box. If they receive a box with less number and or lower quality of books, they're more likely to cancel the box.

- On the other hand, the number of crafts or other items in the box doesn't help retention. In fact, the target audience cares about the books more than the bonus items in their box.

- Customers who subscribed using a coupon or multiple coupons are more price sensitive and more likely to cancel.

- VIP customers who prepaid 1 year of the subscription are more loyal and less likely to cancel.

- Girls like our box better than boys probably because girls are early in developing language skills and or the box is lacking activities that boys like.

- Customers who live in the midwest states like Texas, Ohio and Michigan and show higher mean and median revenues than the west coast and east coast probably due to less access to alternatives and thus our products add more value to these demographics.

With these insights learned from the analysis, we can consider to apply the following practices into the business:

1) Focus on the books selection and quality while keep the extracurricular activities in a separate product line targeting specific groups of customers;

2) Increase the benefits of VIP members to generate more prepaid 1 year subscriptions but also provide a fair amount of basic features to month to month subscribers.

3) Use coupons smartly - free box promotion generates a lot of traffic but leads to high churn rate and overall lower profits.

4) For existing customers with high probability to cancel, try to reach out proactively to learn feedback and find solutions to increase retention.

5) Focus the business development efforts in the midwest states and or English-speaking only families.

Problem Statement

As one of the confounders of a subscription box small business, I am interested in learning from data to get insights about customer behaviors. Specifically, I would like to study the impact of customer demographics, marketing campaigns (including coupons, email interactions, product feedback), product content and cost on the growth and retention of customers.

Launched in September 2018, the subscription box delivers high-quality children's books (in Chinese) to families around the world (mainly in the USA, but also Canada and Malaysia) personalized to children's age and each month with a different fun topic. By February 2020, we have serviced 1317 customers with 7153 shipments. As the owner, I would like to learn perspectives from the data to better serve the customers and grow the business.

Since the launch, there have been 1098 cancellations of the subscription. Why so many cancellations? What caused the cancellations? The goal is to:

1) Identify important factors that predict cancellations;
2) Predict the cancellation probability for existing active subscribers.

Data Wrangling and Exploratory Data Analysis

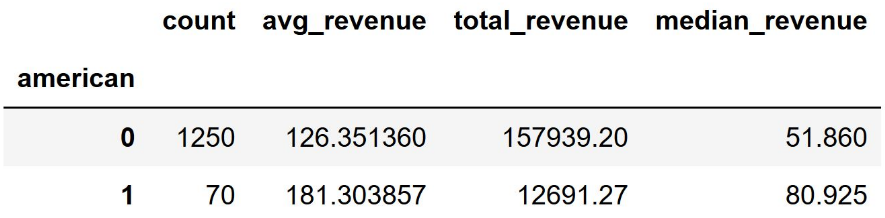
The data are proprietary and were downloaded from the platforms used for the business. Since the data contain sensitive information about customers, the codes, presentations, and slides will not display any individual-level data but only aggregated data.

Customer data (name, ethnicity, address, revenue, num_orders, num_subscriptions, subscription status, payment provider)

There are 1320 customers from 08/2018 to 02/2020. The average revenue generated from each customer is around $130. The highest is $1469 who is a huge fan of our products. The distribution of all revenues is skewed to the right as shown in Figure 1. Interestingly there's a local peak at around 275 which is the annual subscription cost that many customers signed up for.

Table 1 Summarized the total, average and median revenues generated from each state. Intriguingly, midwest states especially TX, OH, MI show higher mean and median revenues than the west and east coasts. This might suggest that families who live in these areas have fewer alternative resources to buy Chinese books for children.

To capture the demographics of customers, a new feature called "American" is created to capture families who don't speak Chinese but want their kids to learn the language. Interestingly, this group generates higher revenue on average and by median. They only represent 5% of the total sample currently and this might suggest we could target the group in the future by launching our English website and bilingual books subscription. A T-Test failed to reject the null hypothesis at a 5% significance level which is there's no difference between the two groups (but was really close with p-value 0.059). Given the small sample, it's worth rerunning the tests when more data is accumulated.



	count	avg_revenue	total_revenue	median_revenue
american				
0	1250	126.351360	157939.20	51.860
1	70	181.303857	12691.27	80.925

Subscription data (customer name, address, subscription status, revenue, subscription start, and end dates, subscription product, subscription term, coupon, is_gift, cancel date)

This is the most important data since our goal is to predict cancellation. There are 309 active, 894 canceled and 71 expired subscribers. The expired group contains one time gifts and customers' credit card expired.

Cancellations due to accidentally duplicated order, VIP testing, and upgrading (they canceled the monthly subscription and upgraded to the annual plan) are excluded from the analysis so only voluntary normal cancellations are studied.

In the raw data, all dates are in PST and therefore are converted to the local time so analysis can be run on the subscription and cancellation date and time. And a new feature is built to capture the lifespan for each customer that starts from their subscription time and cancellation time (if they didn't cancel, the end time is current). Figure 2 and 3 compare the subscribe date and time vs. cancel date and time. It's very interesting to see the cancellation mostly nests around 17th which is the subscription renewal date while the subscription is concentrated 1 week before the new box subscription cutoff date. The time also exhibits some fun findings: both subscribe and cancel happen often from 9 to 10 pm (when the kids are asleep), subscription happens more often in the afternoon while cancelation more in the morning. This might be due to the timing of customers receiving notifications. The system generated notifications are usually sent out late at night while the marketing campaign emails were sent usually in the morning.

Another interesting exploratory data analysis is the total revenue from families with girls is higher than families with boys in terms of mean, median, 25th and 75th percentiles (see Figure 4). A T-test rejected the null hypothesis with a 1% significance level. The story might be girls are earlier than boys in developing languages and thus are more likely to like reading books than boys. The insights we can learn from here is boxes need to be personalized to add items boys like for example cars, crafts, etc.

Figure 5 shows the distribution of the children's ages when they start the subscription and caution needs to be taken here when drawing conclusions. But still, this suggests at what age range the families are more attracted to our product and it's around 2-3 years old.

On top of the coupons used, a couple of new features are created to capture if a customer used coupons to subscribe and how many coupons did they use. Figure 6 groups total revenues by the number of coupons used. The goal of coupons is to incur acquisition costs while generating long term revenues. However, this doesn't see to be the case since the more coupons used the less total revenues generated. At the growth phase of the business, this might be fine especially considering the network effects but over the long term, it's hard to grow customers organically solely using coupons.

Subscription cancellation data (customer name and address, subscription product, subscription term, subscription start date, months subscribed, cancel reason, cancel note)

Figure 7 ranks the reasons for cancelling. Most cancellations are not due to dislike but too many books and no room to put more. Unlike food, books are not fast-moving consumer goods. The leader needs to think if a renting business is more viable (like the early stage of Netflix).

Subscription box content data (month, year, topic, age_group, num_of_books, num_of_crafts, num_of_items, costs of the box)

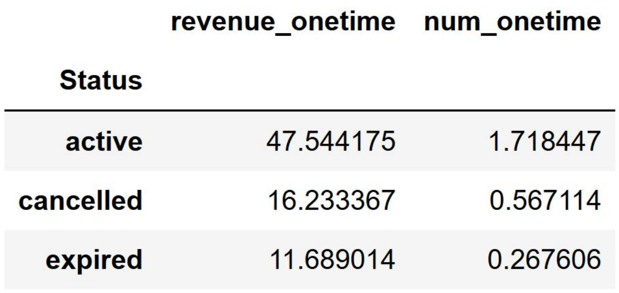
These are the contents for each month's box in terms of how many and what kind of books or other items are. New features are created to capture the exponential weighted average of # of books and other items and costs for the boxes received (before cancellation). So more weights are given to more recent boxes received. This is trying to capture if the content of a box influences the decisions of customers to continue or cancel, especially for the most recent boxes. Figure 8 suggests that fewer books and or lower costs (cheaper books) might explain cancellation. On the other hand, the number of items doesn't seem to matter. After all, the main goal of subscribing to the box is still reading.

Customer feedback data (customer name, shipped date, customer since, rating, response, subscription status)

A new feature is included to measure the length of the feedback. Figure 9 suggests most of the ratings received are above average which is good. Figure 10 is very interesting that the longest feedback happens for ratings of 2 and 3 when customers want to give meaningful feedback and hope they can see improvements later. When customers give a rating of 5 they don't have much to say and when they are very angry and give 1 they don't want to waste their time.

One time purchase data (customer name, purchase date, purchase time, revenue, address, coupon, subscription status, product names)

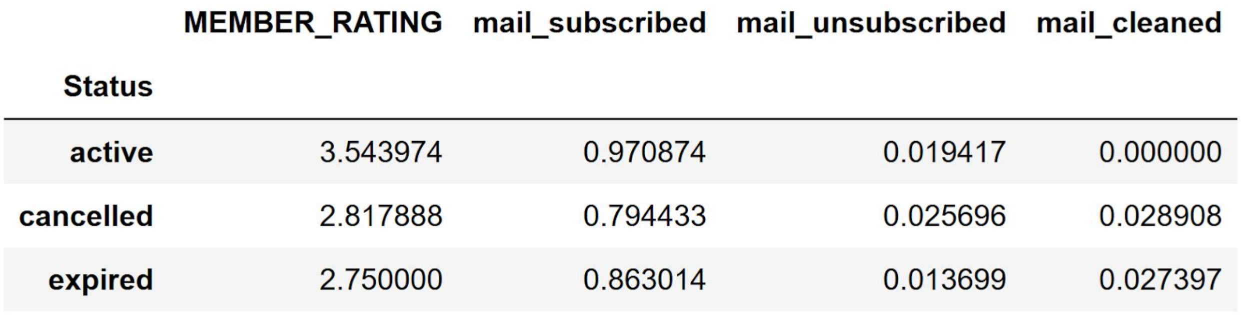
On top of the subscription business, we also have an online shopping page that sells one-time items for children's books. It seems the active subscribers really like books and order more one-time items than the canceled groups.



	revenue_onetime	num_onetime
Status		
active	47.544175	1.718447
cancelled	16.233367	0.567114
expired	11.689014	0.267606

Mailchimp email campaign data (campaign title and subject, send date and time, # of recipients, engagement rating, unsubscribes, cleans)

Each week we send out marketing email campaigns to all email subscribers. The features that could be useful in predicting cancellations are the member engagement rating and if they unsubscribed or cleaned from the email list. Not surprisingly, customers with higher email engagement ratings and doesn't unsubscribe to email lists are less likely to cancel their subscription.



	MEMBER_RATING	mail_subscribed	mail_unsubscribed	mail_cleaned
Status				
active	3.543974	0.970874	0.019417	0.000000
cancelled	2.817888	0.794433	0.025696	0.028908
expired	2.750000	0.863014	0.013699	0.027397

In-depth Analysis and Machine Learning

Goal

The goal of the in-depth analysis is, given the features of customer demographics (child age, language, gender), subscription age and term, coupon used, number of items and costs for the boxes received, feedback rating, and email interaction score to predict cancellation.

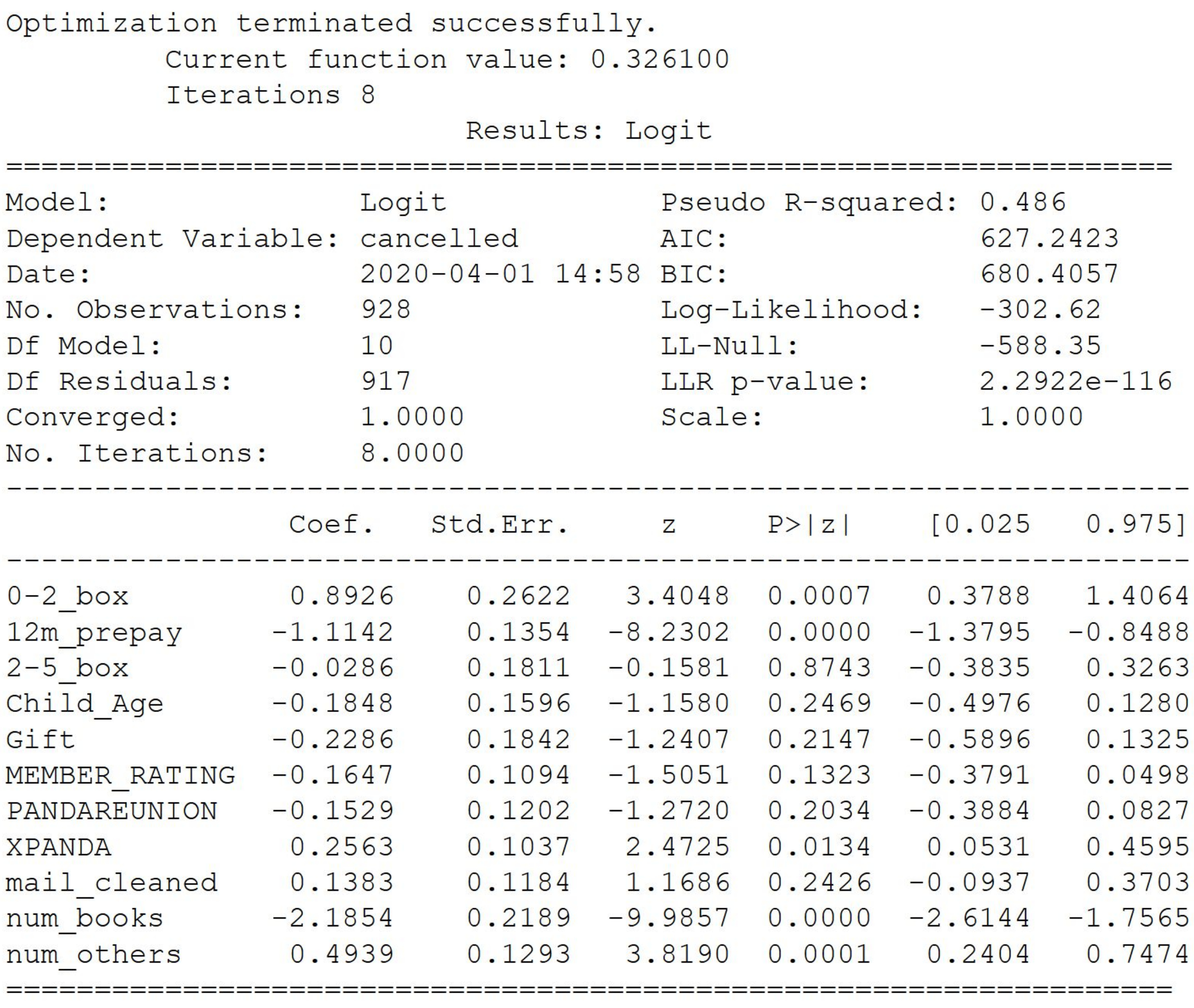
Data Preparation

Training data and testing data were split from the original data with 70% and 30% respectively. Given the list of features: child age, main language speaking, gender, subscription box age and term, name of coupons used and # of coupons used, # of books and other items and costs of boxes received, feedback rating, email interaction score and if unsubscribed from the mailing list.

There were a couple of months when we tried out a very aggressive campaign and people could try out the box for free in the first month. The campaign was not very successful in terms of high volumes of cancellations happening before they received the box. Since we know the reasons for these types of cancellations, these observations were excluded. You can see this in Figure 11. There are 649 observations in the training data and 279 in the testing data with 74 features.

Feature Selection

Since we have 74 features with many of them sparse data, recursive feature elimination (RFE) was used for feature selections. It's based on the idea to repeatedly construct a model to choose the best features recursively. Features are selected by the ranking and the correlation matrix (Figure 12) to run a first round of logistic regression.



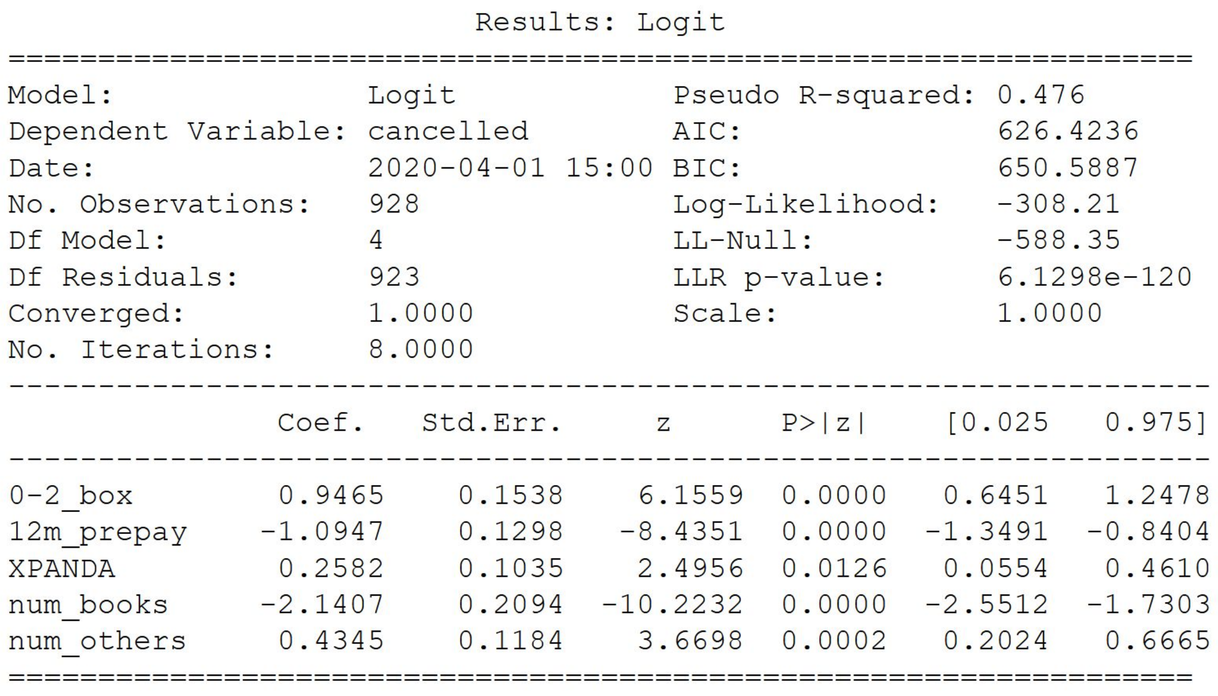
```
Optimization terminated successfully.
         Current function value: 0.326100
         Iterations 8
                          Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared: 0.486
Dependent Variable: cancelled        AIC:              627.2423
Date:               2020-04-01 14:58 BIC:              680.4057
No. Observations:   928              Log-Likelihood:   -302.62
Df Model:           10               LL-Null:          -588.35
Df Residuals:       917              LLR p-value:      2.2922e-116
Converged:          1.0000           Scale:            1.0000
No. Iterations:     8.0000
-------------------------------------------------------------------
                Coef.   Std.Err.     z     P>|z|    [0.025   0.975]
-------------------------------------------------------------------
0-2_box         0.8926    0.2622   3.4048  0.0007   0.3788   1.4064
12m_prepay     -1.1142    0.1354  -8.2302  0.0000  -1.3795  -0.8488
2-5_box        -0.0286    0.1811  -0.1581  0.8743  -0.3835   0.3263
Child_Age      -0.1848    0.1596  -1.1580  0.2469  -0.4976   0.1280
Gift           -0.2286    0.1842  -1.2407  0.2147  -0.5896   0.1325
MEMBER_RATING  -0.1647    0.1094  -1.5051  0.1323  -0.3791   0.0498
PANDAREUNION   -0.1529    0.1202  -1.2720  0.2034  -0.3884   0.0827
XPANDA          0.2563    0.1037   2.4725  0.0134   0.0531   0.4595
mail_cleaned    0.1383    0.1184   1.1686  0.2426  -0.0937   0.3703
num_books      -2.1854    0.2189  -9.9857  0.0000  -2.6144  -1.7565
num_others      0.4939    0.1293   3.8190  0.0001   0.2404   0.7474
==================================================================
```

Logistic Regression

The list of features is further reduced by removing insignificant variables (p-value > 0.05). And so we have the final list of features as: 0-2 box, 12m_prepay, XPANDA (which is 40% off the 1st box), # of books, and # of others used in another round of logistic regression.

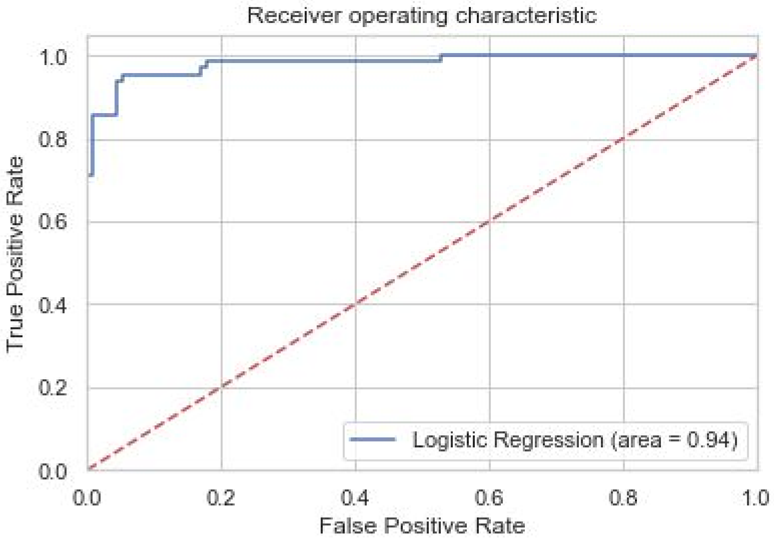
The coefficients are reasonable:

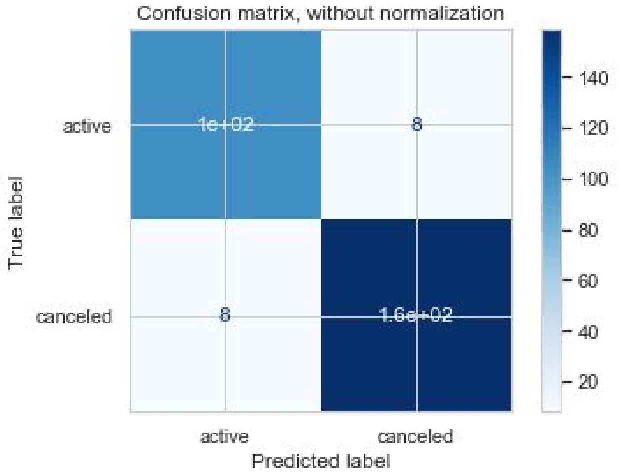
- 0-2 box subscribers are more likely to cancel probably because new moms are enthusiastic about trying the products but realize their babies are still way too young to start the reading journey.
- VIP subscribers are less likely to cancel - when customers prepay the 1 year subscription they’re more likely to be our fans.
- If customers used the coupon XPANDA and took 40% off the 1st box, they’re more likely to cancel, reflecting their price sensitiveness to the product.
- If there are more books and less other items in the box, they’re less likely to cancel, suggesting we should focus on the core of the box which is book selection and quality.

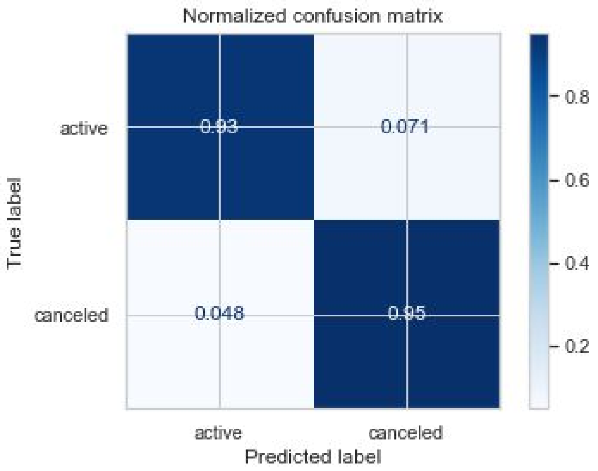


```
                          Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.476
Dependent Variable: cancelled        AIC:              626.4236
Date:               2020-04-01 15:00 BIC:              650.5887
No. Observations:   928              Log-Likelihood:   -308.21
Df Model:           4                LL-Null:          -588.35
Df Residuals:       923              LLR p-value:      6.1298e-120
Converged:          1.0000           Scale:            1.0000
No. Iterations:     8.0000
------------------------------------------------------------------
               Coef.   Std.Err.     z      P>|z|    [0.025   0.975]
------------------------------------------------------------------
0-2_box        0.9465    0.1538    6.1559  0.0000   0.6451   1.2478
12m_prepay    -1.0947    0.1298   -8.4351  0.0000  -1.3491  -0.8404
XPANDA         0.2582    0.1035    2.4956  0.0126   0.0554   0.4610
num_books     -2.1407    0.2094  -10.2232  0.0000  -2.5512  -1.7303
num_others     0.4345    0.1184    3.6698  0.0002   0.2024   0.6665
=================================================================
```

This model leads to AUC of 0.94, accuracy score of 0.94, precision 0.95, recall 0.95 and F1 score of 0.95. Not bad at all!

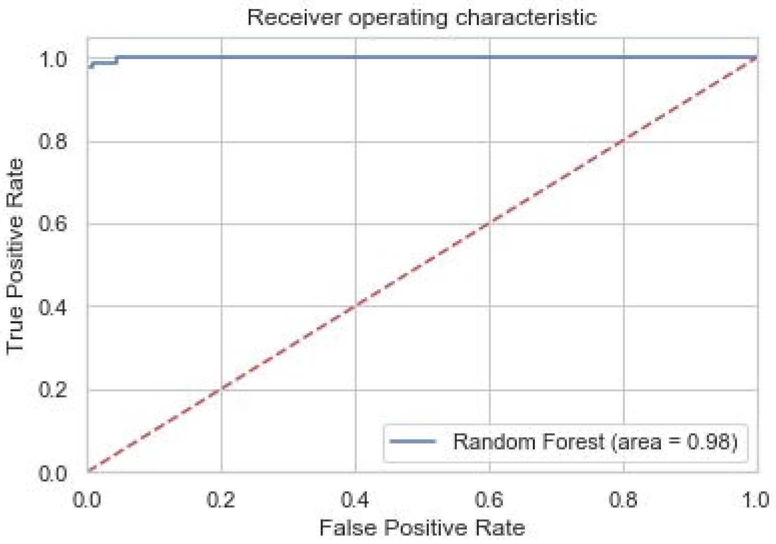


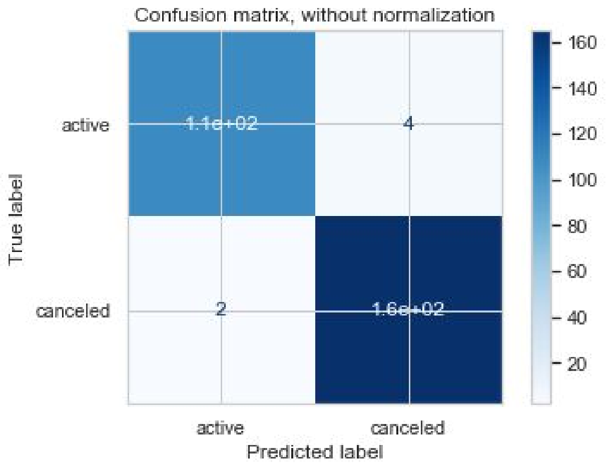


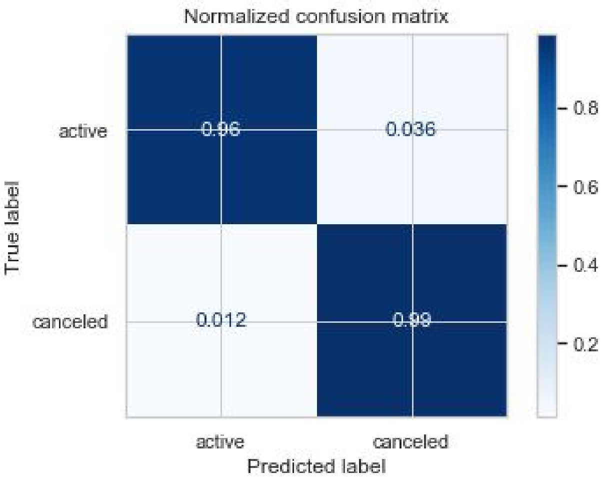


Random Forest

A random forest model with n_estimators=100 and max_depth=3 was tested and the feature importances is seen in Figure 13. AUC is 0.98 with accuracy score, precision, recall, and F1 score all are around 0.98. That’s pretty good!







Caveat
The number of observations is quite small. More data should be accumulated and the analysis needs to be rerun.

There are apparently some important missing features for example customers' logging activities , social media referrals and engagements which are hard to collect with the current platform but could be studied in the future.

Conclusions

The prediction model performs extremely well and can be used by the customer experience team to proactively reach out to existing subscribers who have high cancellation probability and gather feedback and provide solutions.

The insights learned from the significant features that predict cancellations are tremendously important. Based on the findings, the product team needs to focus on book selection and quality while expanding new product lines for age 0-2, boys, and families who only speak English. The technology team should put efforts in building the website with an English version to attract more English speaking only families. The marketing team needs to be more creative than simply providing price discount promotions since those are less likely to build loyal customers. They should also think of ways to add more benefits than simply the boxes such as free webinars, education contents. The sales team needs to put more resources in midwest states where customers have less access to quality Chinese books.

Figure 1.

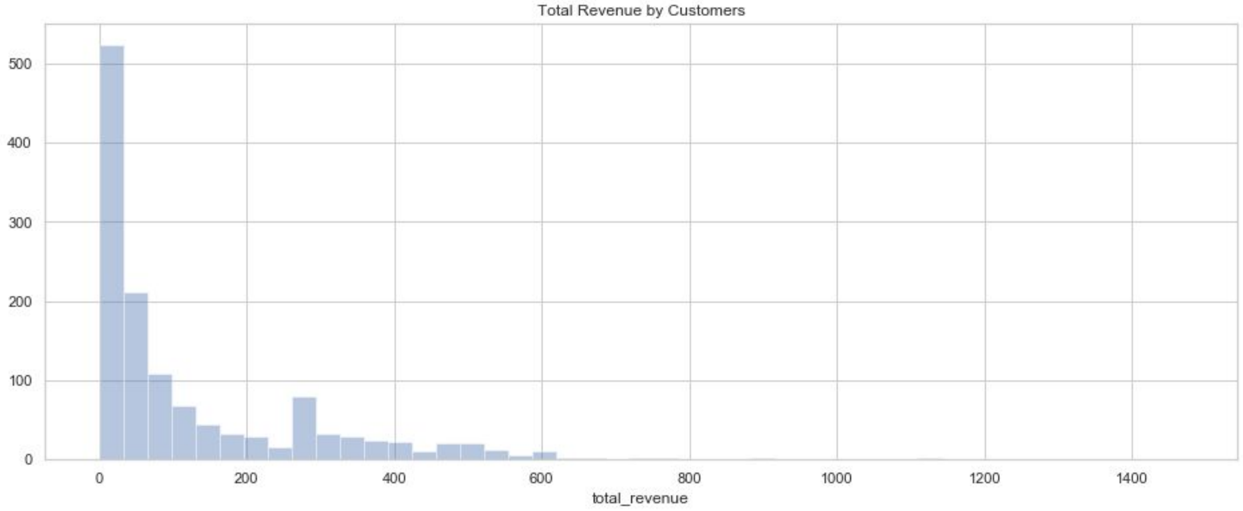
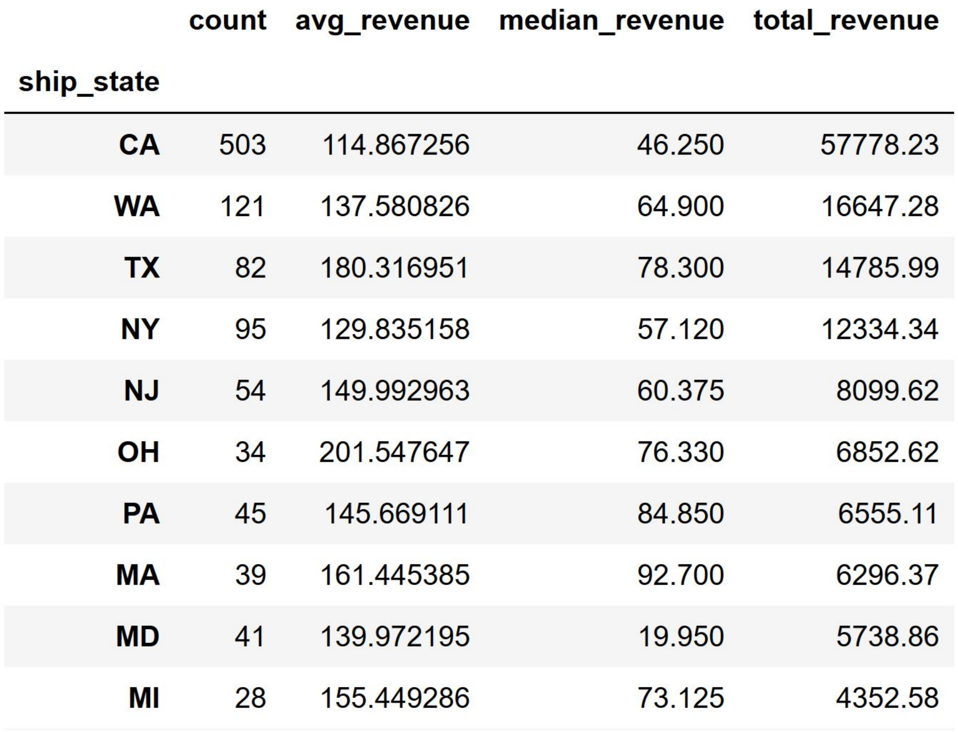


Table 1.



ship_state	count	avg_revenue	median_revenue	total_revenue
CA	503	114.867256	46.250	57778.23
WA	121	137.580826	64.900	16647.28
TX	82	180.316951	78.300	14785.99
NY	95	129.835158	57.120	12334.34
NJ	54	149.992963	60.375	8099.62
OH	34	201.547647	76.330	6852.62
PA	45	145.669111	84.850	6555.11
MA	39	161.445385	92.700	6296.37
MD	41	139.972195	19.950	5738.86
MI	28	155.449286	73.125	4352.58

Figure 2.

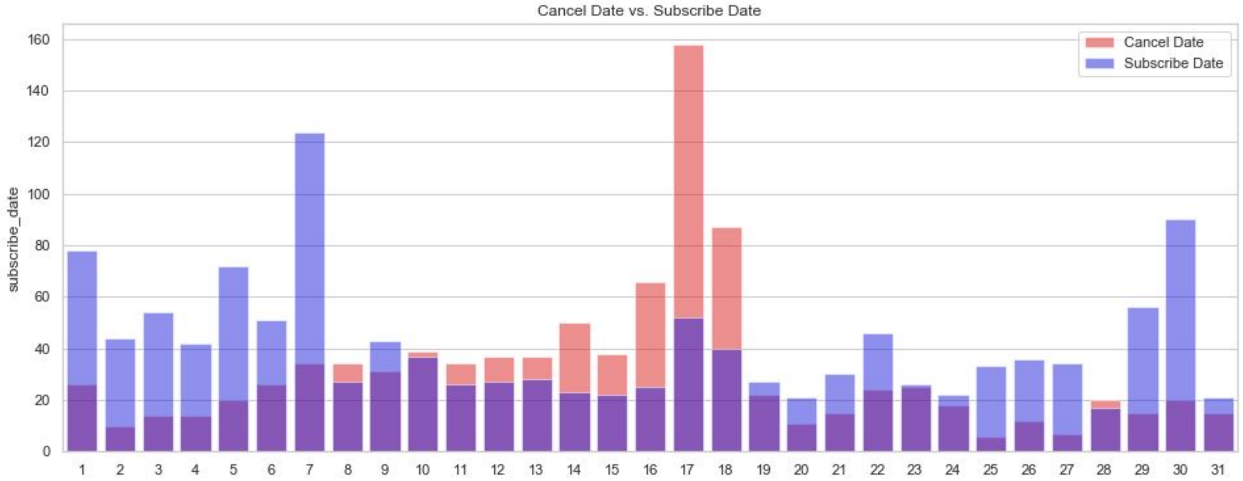


Figure 3.

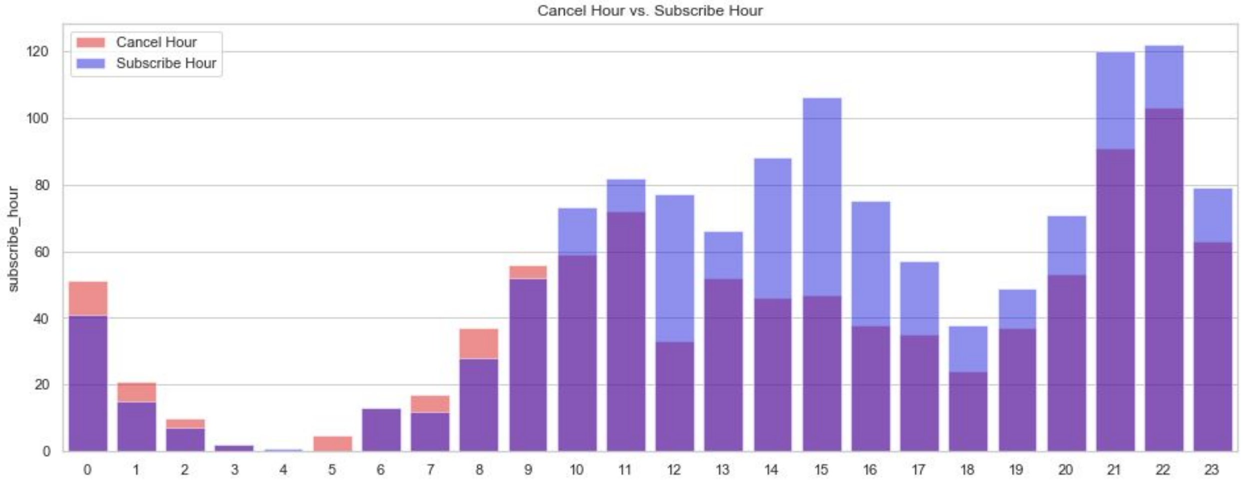


Figure 4.

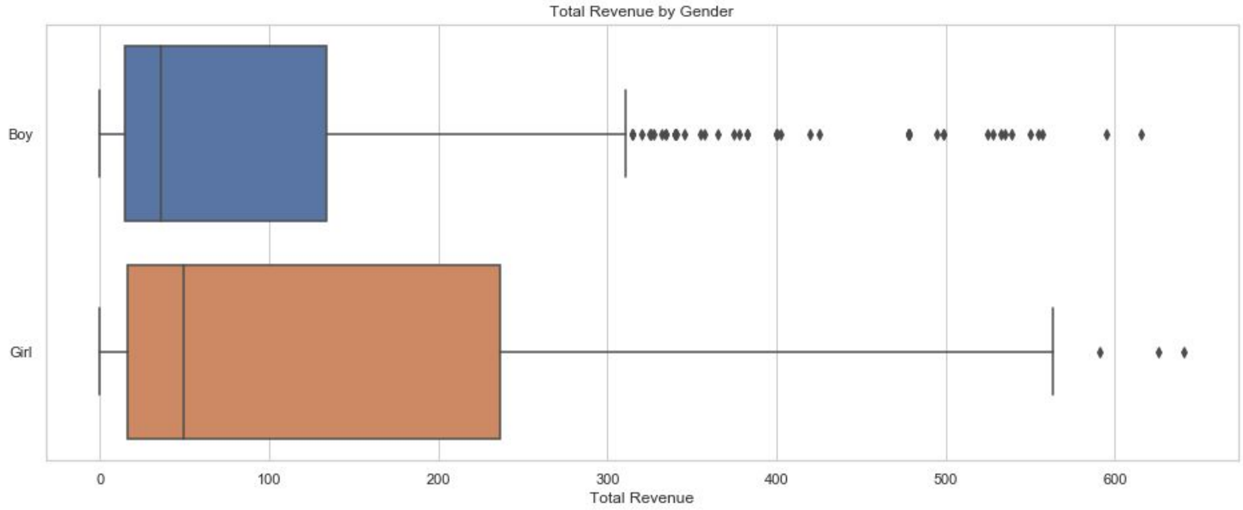


Figure 5.

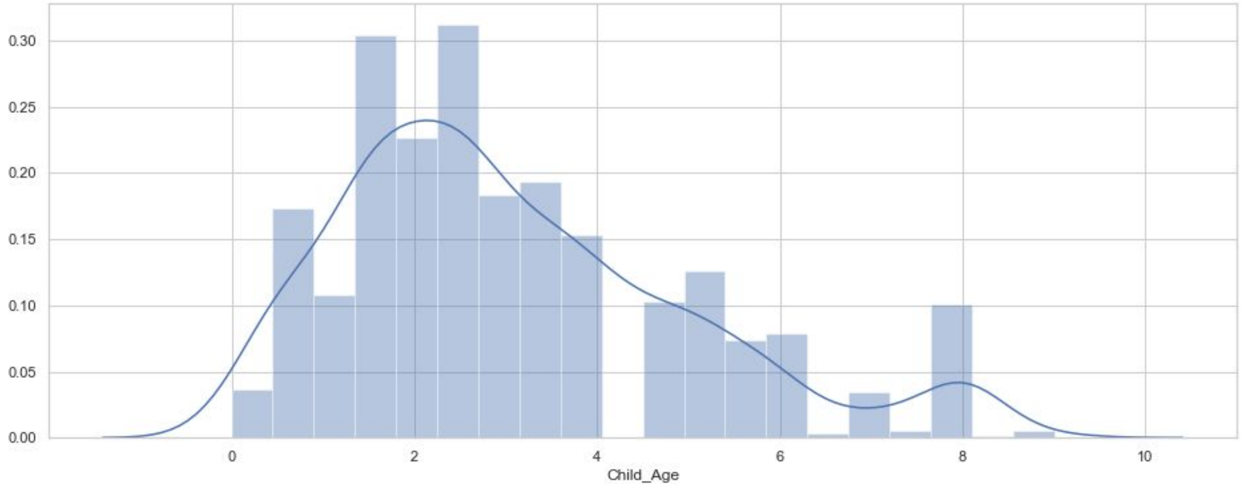


Figure 6.

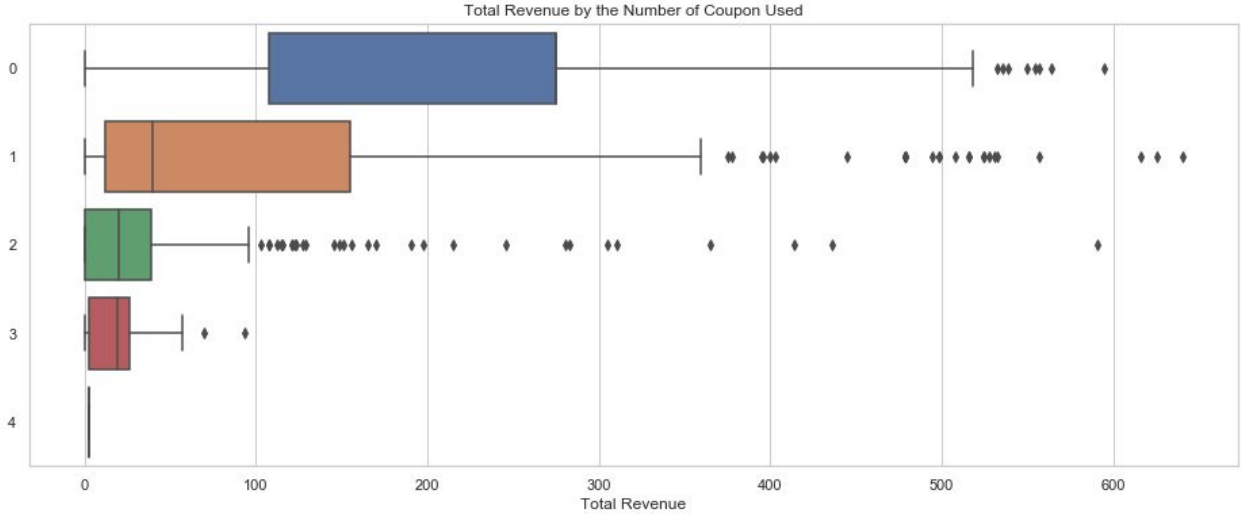


Figure 7.

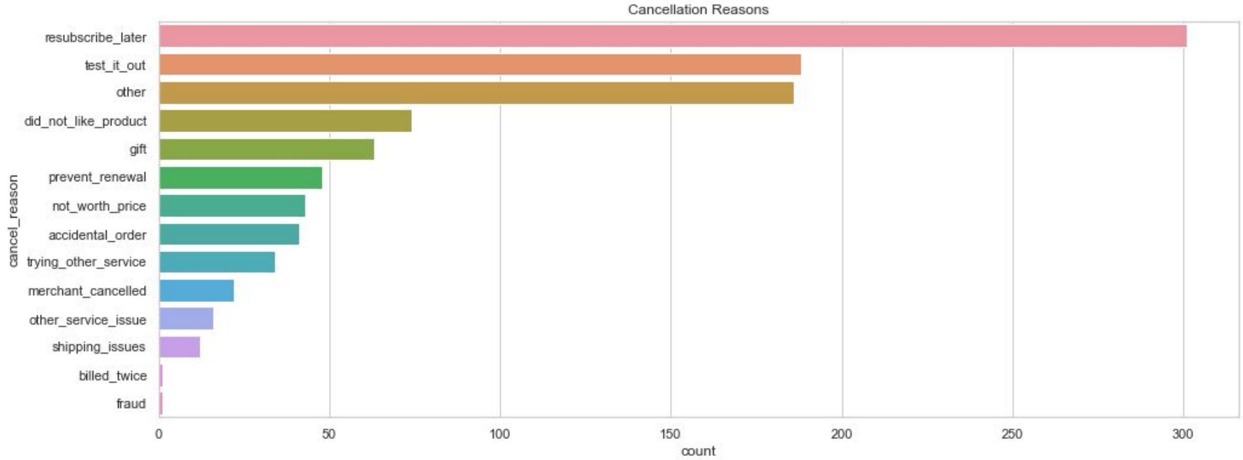


Figure 8.

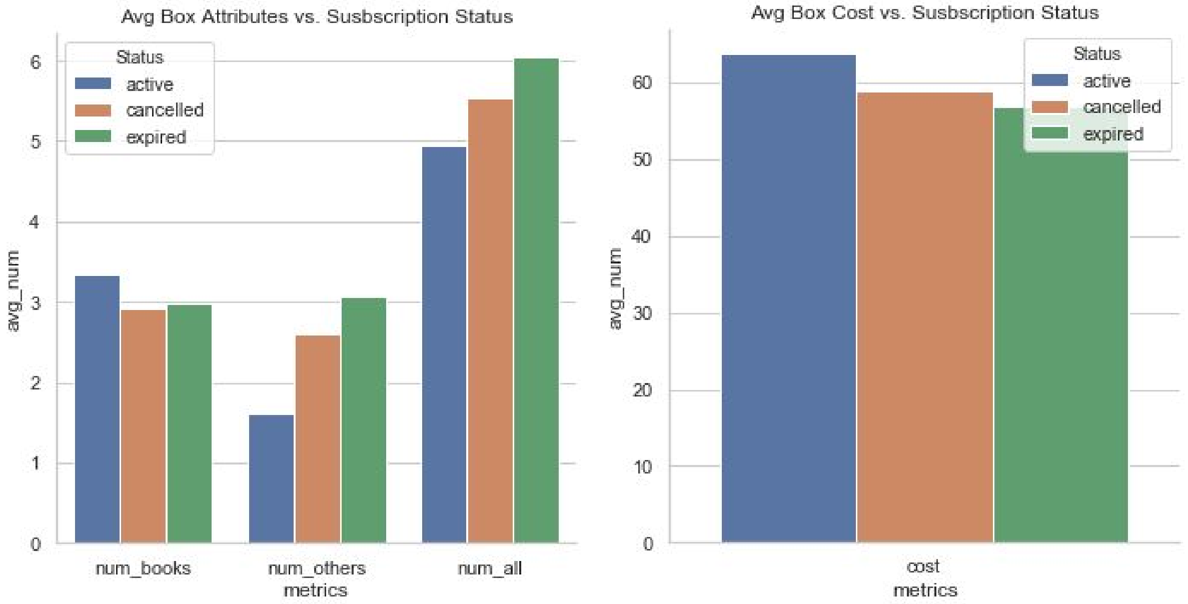


Figure 9.

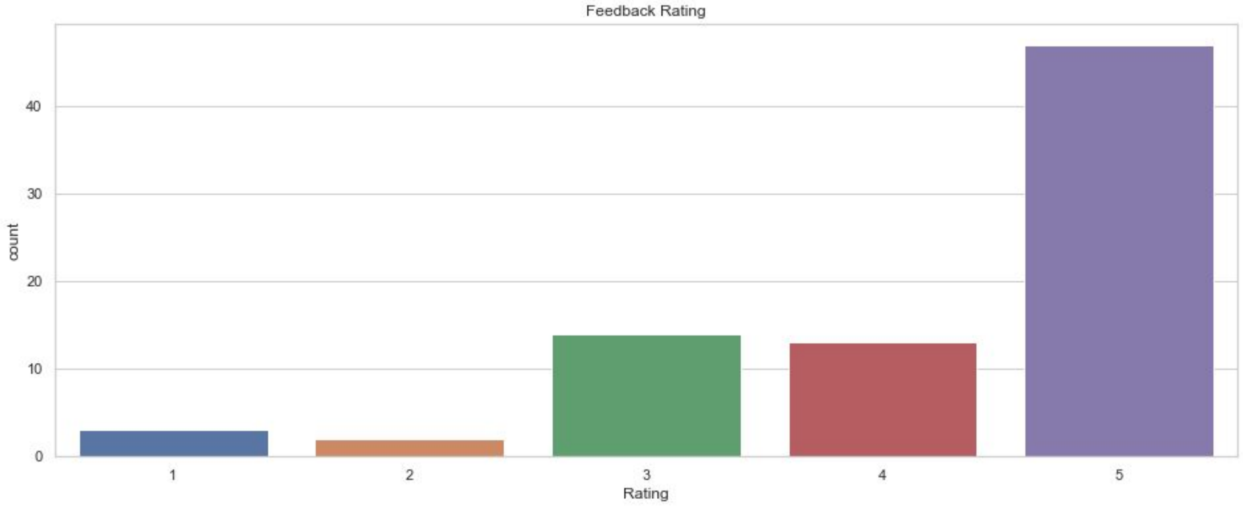


Figure 10.

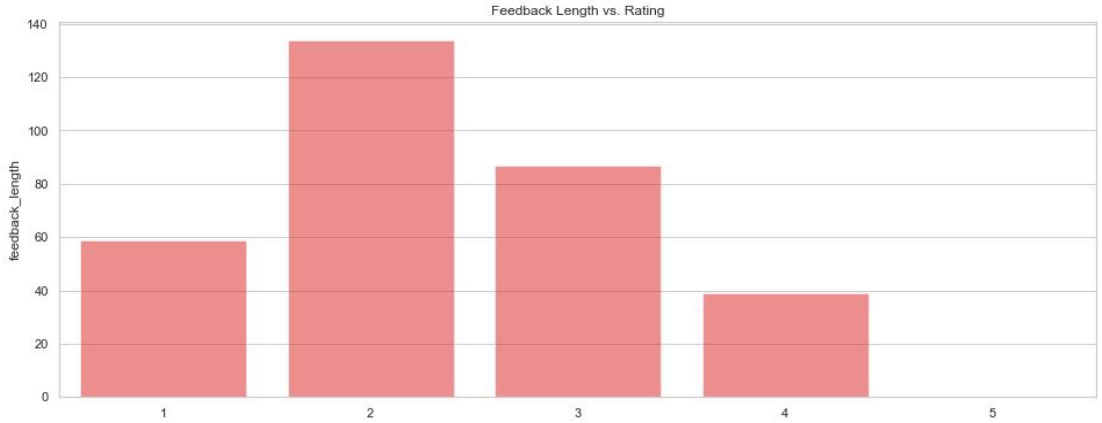


Figure 11.

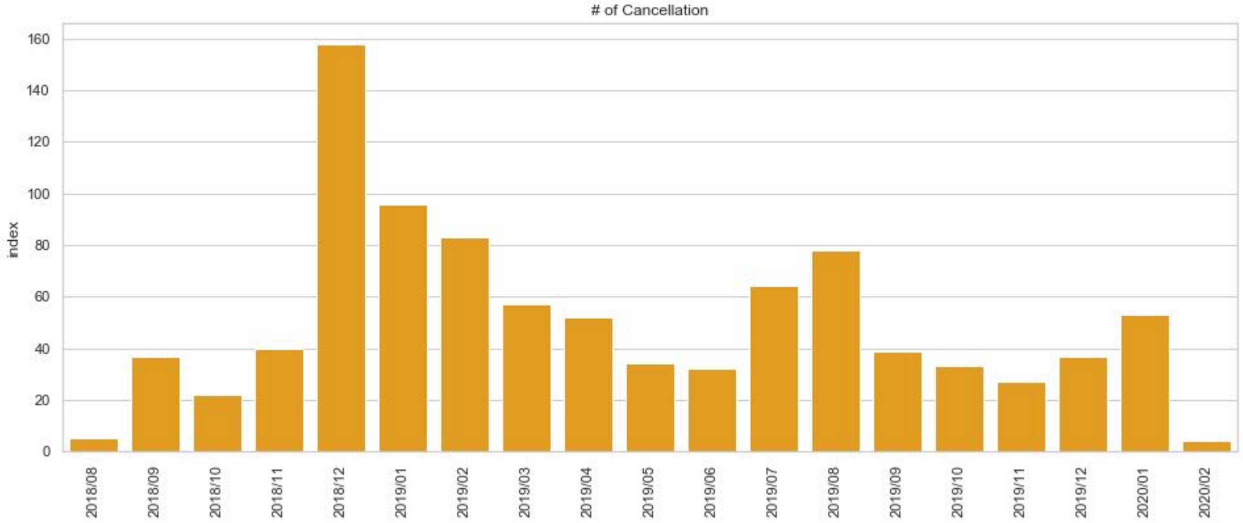


Figure 12.

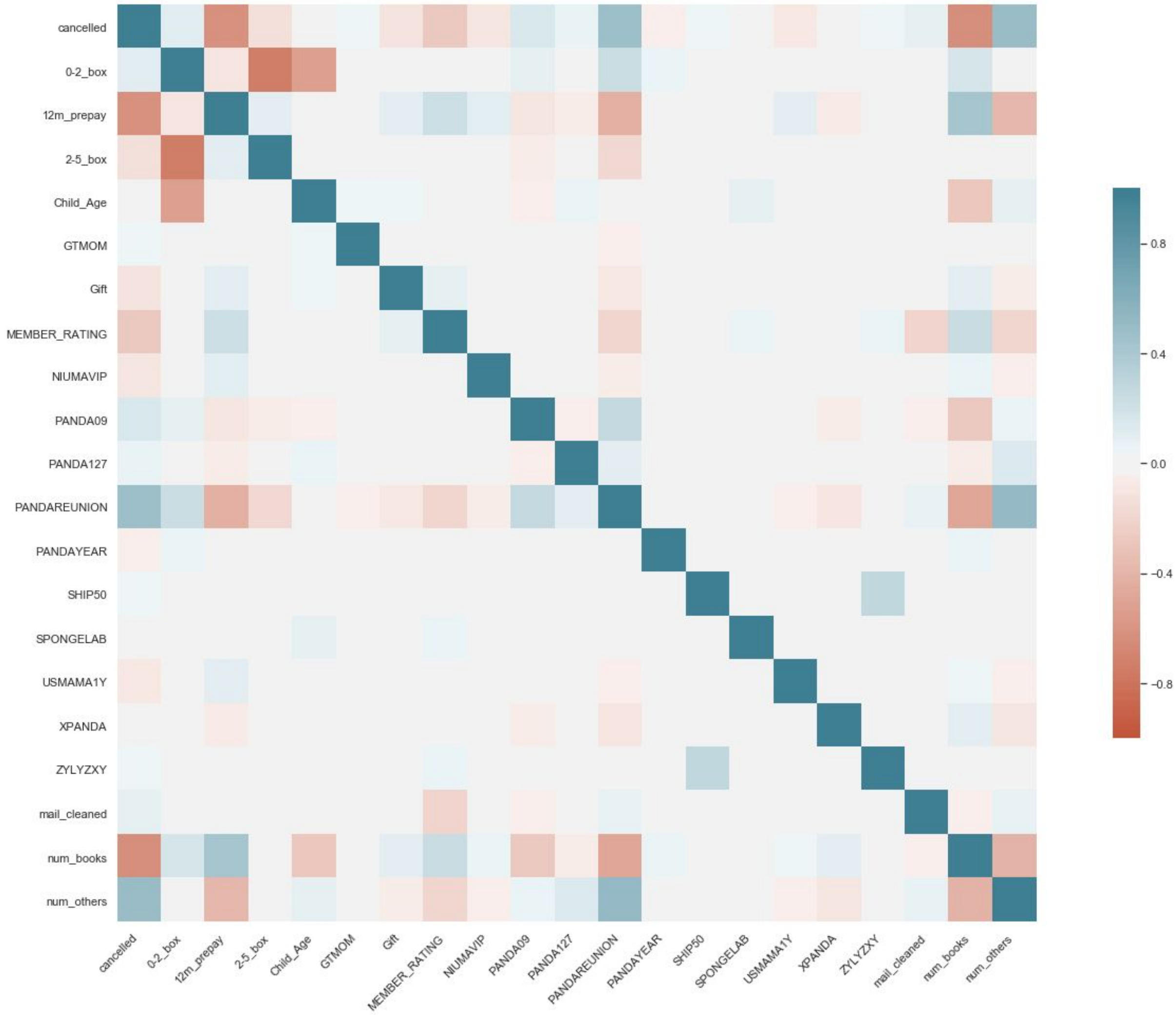
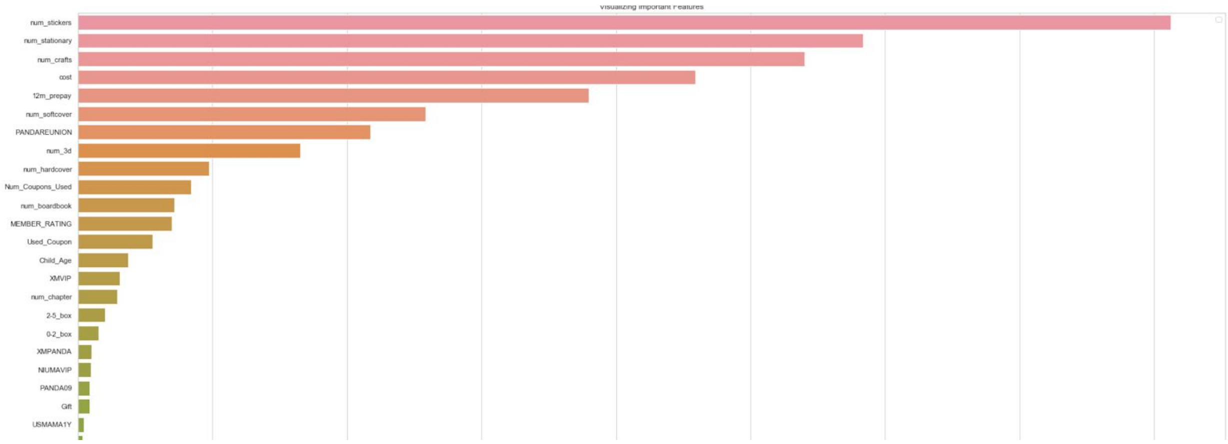

cancelled
0-2_box
12m_prepay
2-5_box
Child_Age
GTMOM
Gift
MEMBER_RATING
NIUMAVIP
PANDA09
PANDA127
PANDAREUNION
PANDAYEAR
SHIP50
SPONGELAB
USMAMA1Y
XPANDA
ZYLYZXY
mail_cleaned
num_books
num_others
0.8
0.4
0.0
−0.4
−0.8

Figure 13.


Visualizing Important Features
num_stickers
num_stationary
num_crafts
cost
12m_prepay
num_softcover
PANDAREUNION
num_3d
num_hardcover
Num_Coupons_Used
num_boardbook
MEMBER_RATING
Used_Coupon
Child_Age
XMVIP
num_chapter
2-5_box
0-2_box
XMPANDA
NIUMAVIP
PANDA09
Gift
USMAMA1Y