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.

- 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.
- 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.

Other Fun Insights

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

Problem Statement

Why customers cancelled?

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.

Goal

Why customers cancelled?

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

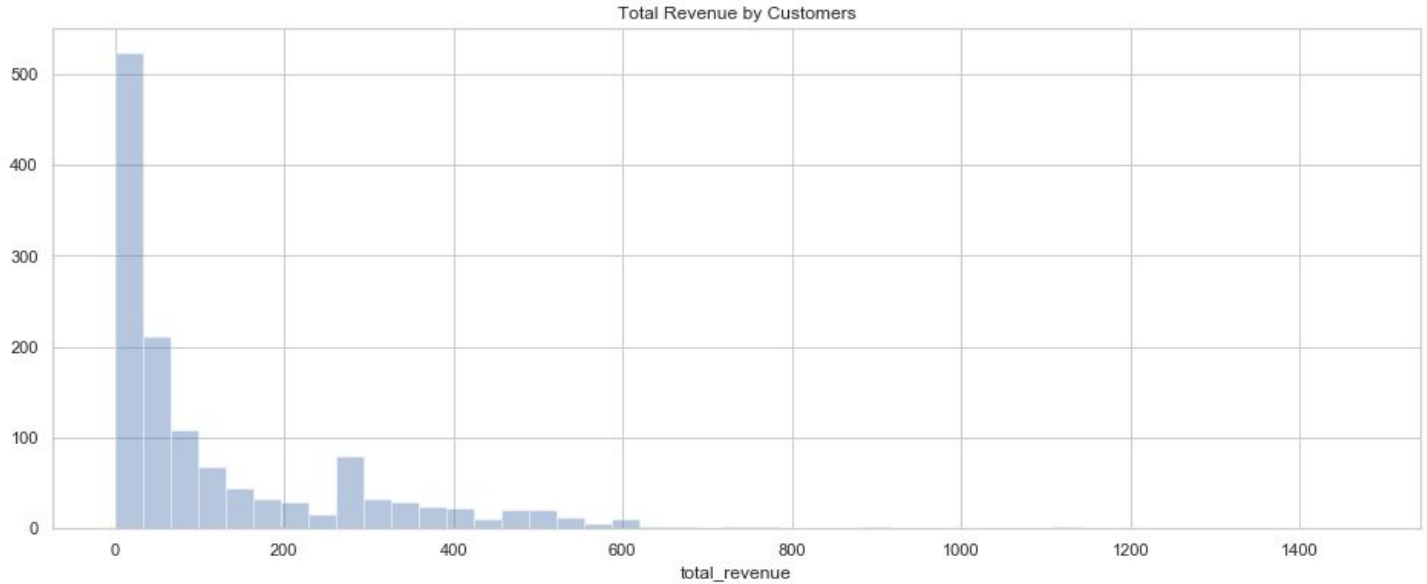
1) Identify important factors that predict cancellations
2) Predict the cancellation probability for existing active subscribers

Customer Data

- There are 1320 customers from 08/2018 to 02/2020. The average revenue generated from each customer is around $130.
- Midwest states especially TX, OH, MI show higher mean and median revenues than the west and east coasts.
- Families who only speak English but want their kids to learn Chinese generates higher revenue on average and by median. A T-Test failed to reject the null hypothesis at a 5% significance level but was very close.

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

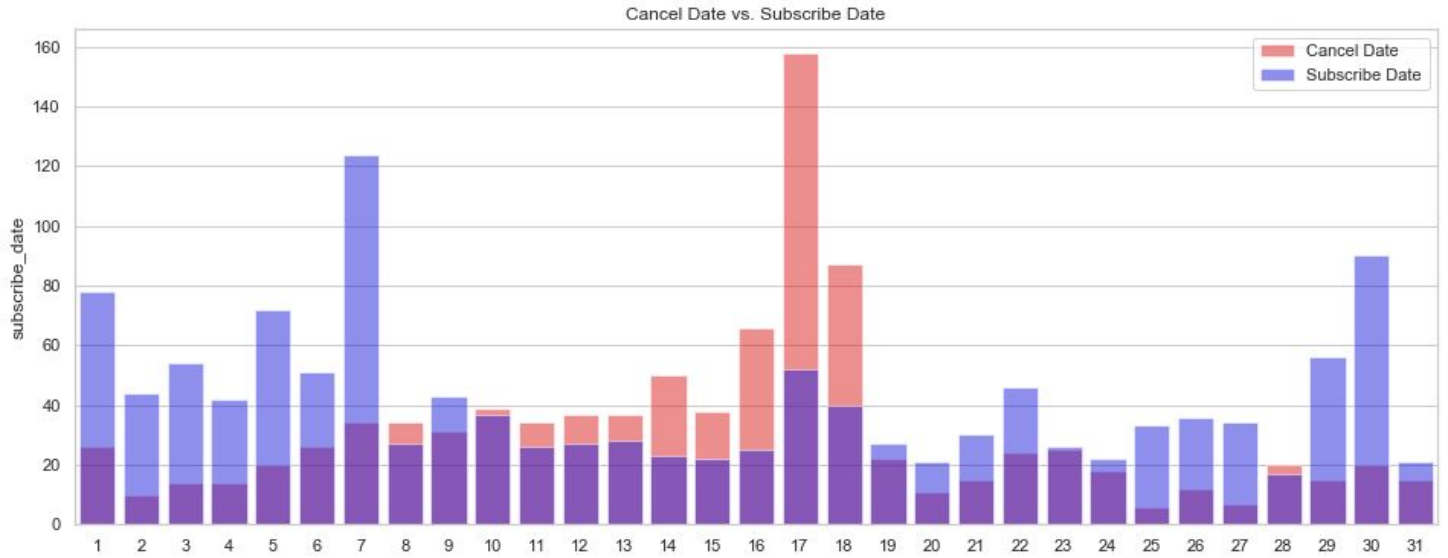
Customer Data Histogram



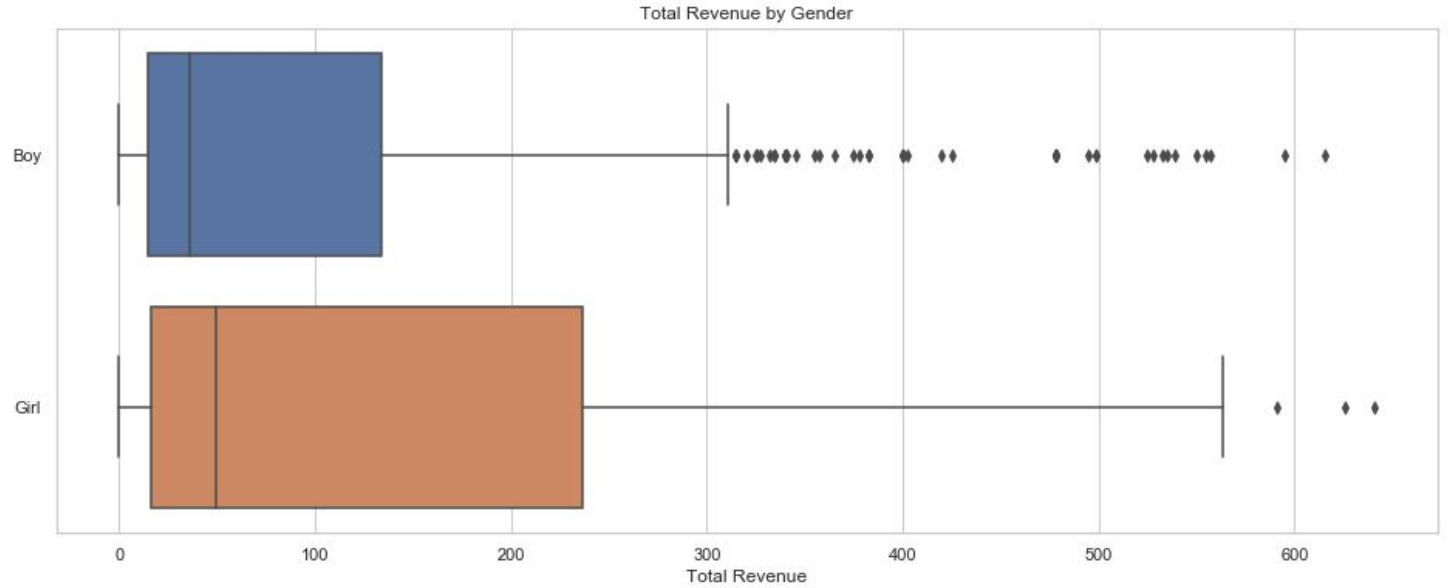
Subscription Data

- 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.
- 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.
- Another interesting exploratory data analysis is the total revenue from families with girls is higher than families with boys. 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.

Cancellation Date vs. Subscription Date

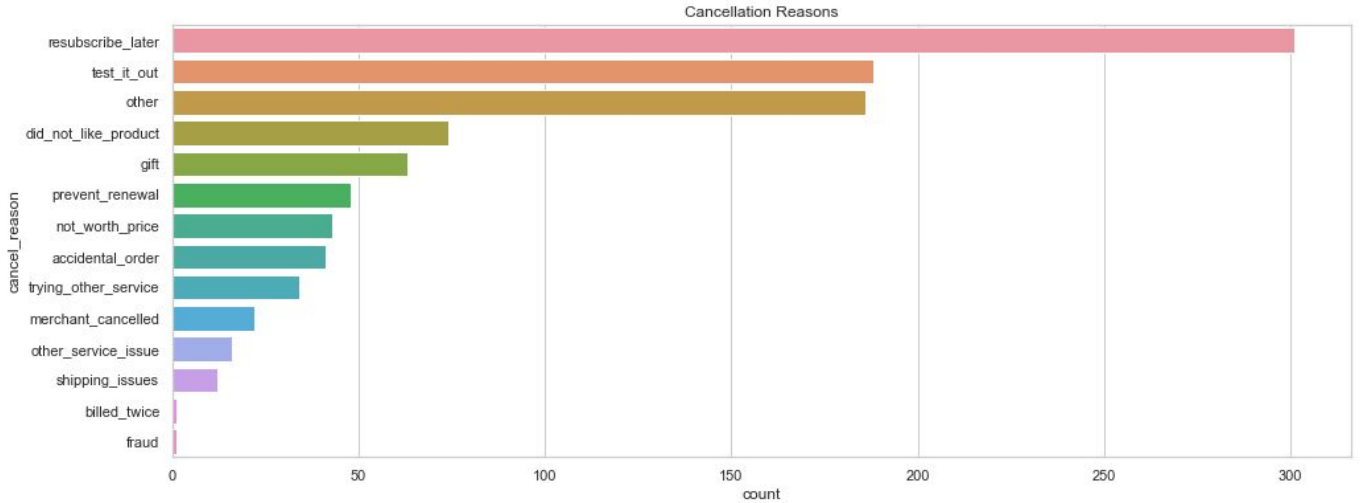


Girls vs. Boys



Subscription Cancellation Reasons

- 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.

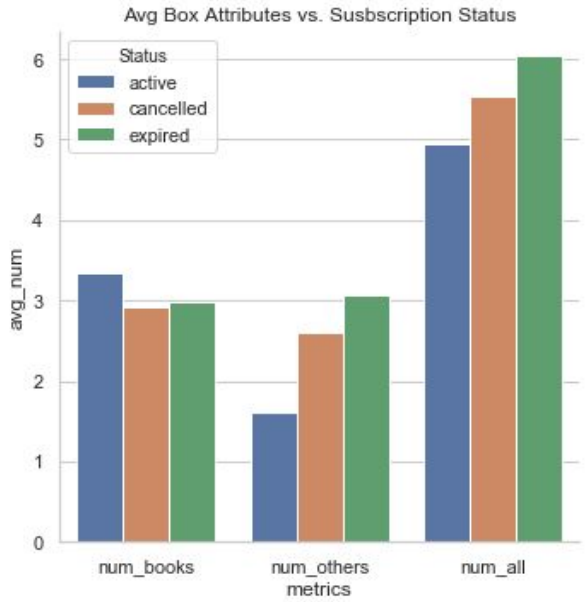


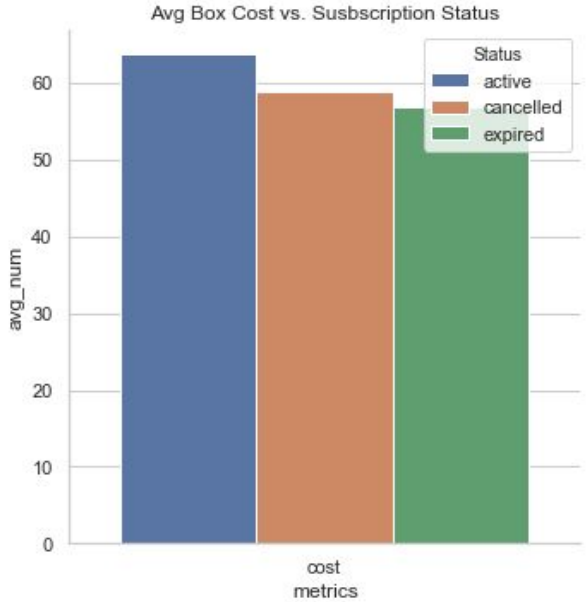
Subscription Box Content Data

- 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.

- 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.

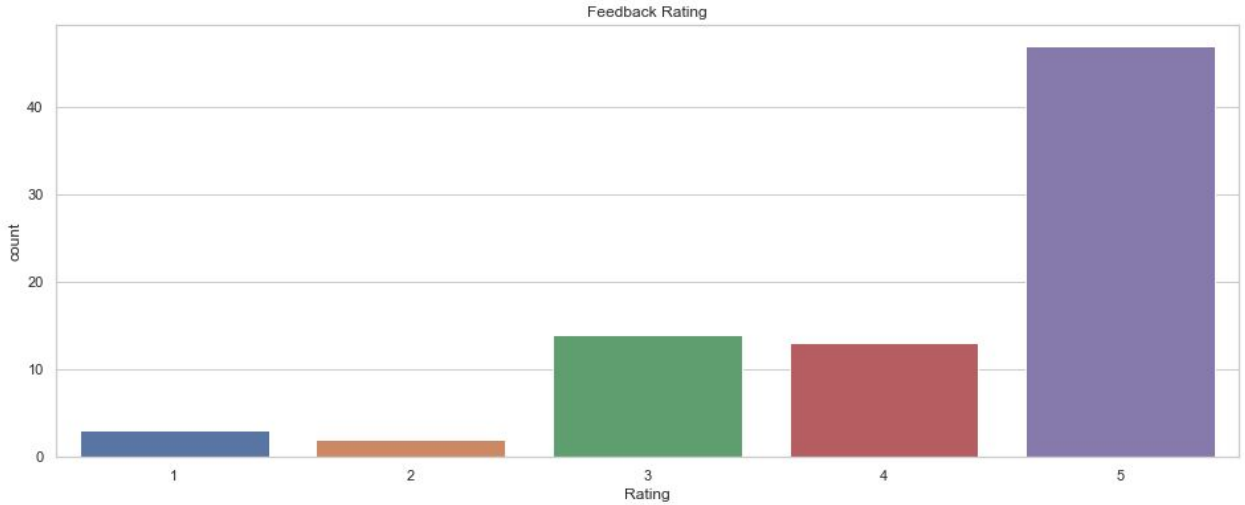
Book Quantity and Quality Matter





Customer Feedback Data

- Most of the ratings received are above average. Interestingly, the longest feedback happens for ratings of 2 and 3 when customers want to give meaningful feedback and hope they can see improvements later.



One-time Purchase Data

- 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.

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

Email Campaign Data

- 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. Customers with higher email engagement ratings and doesn't unsubscribe to email lists are less likely to cancel their subscription.

Status	MEMBER_RATING	mail_subscribed	mail_unsubscribed	mail_cleaned
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

- The goal of the in-depth analysis is, 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 to predict cancellation.
- Training data and testing data were split from the original data with 70% and 30% respectively.
- 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.
- There are 649 observations in the training data and 279 in the testing data with 74 features.

Feature Selection

- Recursive feature elimination (RFE) was used for the first round of feature selections.

	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

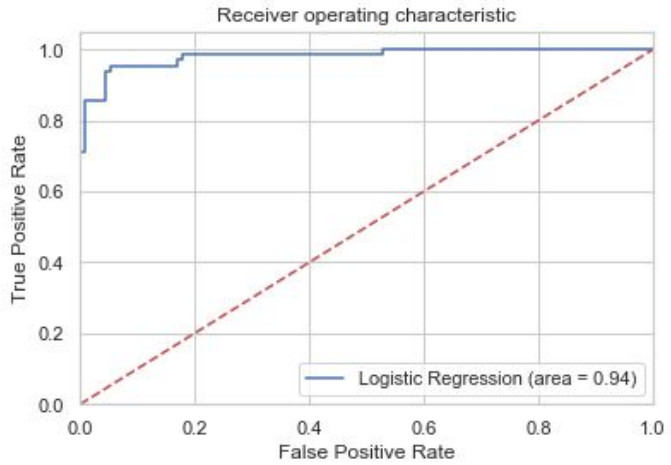
Logistic Regression

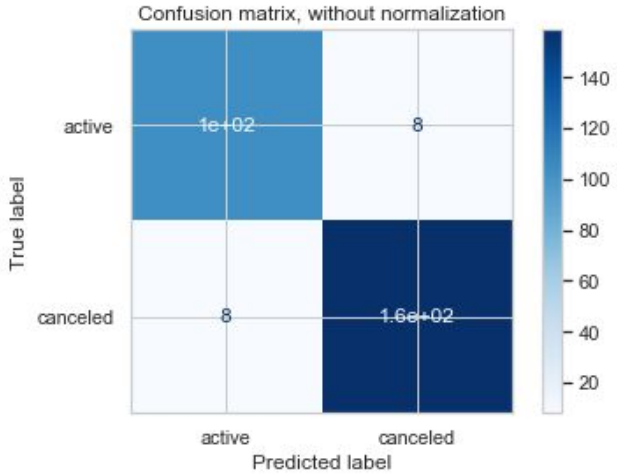
- The list of features is further reduced by removing insignificant variables (p-value > 0.05). And so we have the final list of features as: 0-2 box, 12m_prepay, XPANDA (which is 40% off the 1st box), # of books, and # of others used in another round of logistic regression.

	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

Logistic Regression Model Performance

- 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!





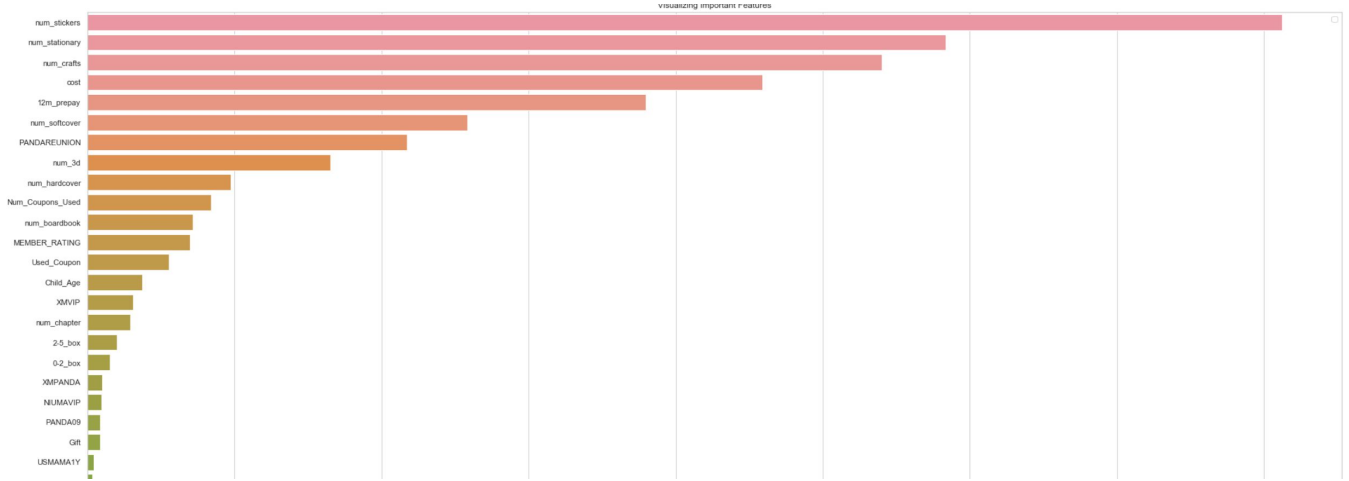
Coefficient Interpretation

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.

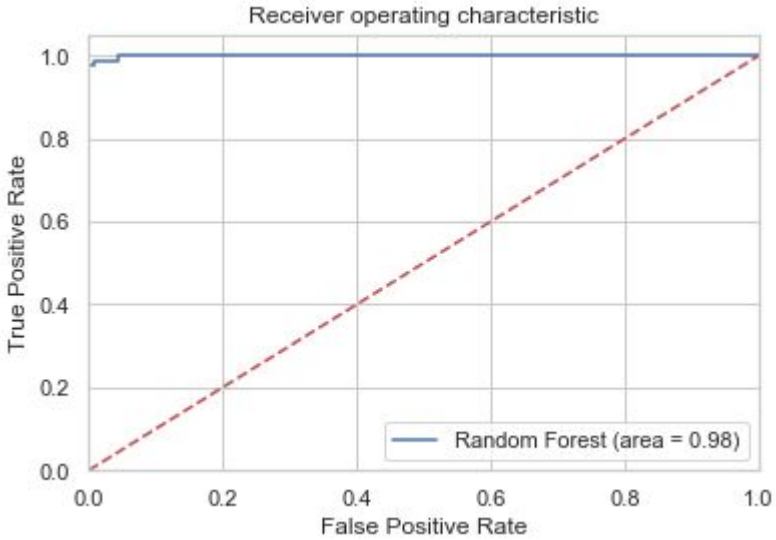
Random Forest

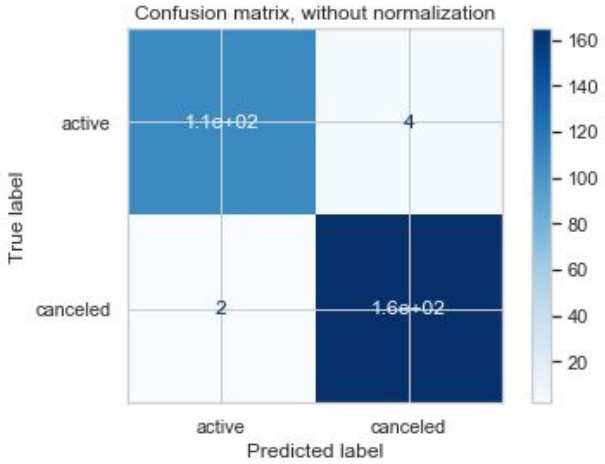
- A random forest model with n_estimators=100 and max_depth=3 was tested and the feature importances are ranked as follows:



Random Forest Model Performance

- 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.

Practical Takeaways

The insights learned from the significant features that predict cancellations are tremendously important.

- 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.