MA 678 Midterm Project

Kunyu Liu

1. Abstract

As the convenience of shopping online nowadays, shopping online becomes more and more popular. Unlike shopping in-store, shopping online has unique features. In this project, statistical learning methods were applied to find attribution affect the success of a product in online shopping and Bayesian regression is used to access the correlation between these variables and the success of a product. Our results indicate that there are some factors that have a strong correlation with product success, but there also exist some limitations. Additional research is needed in this project to develop comprehensive results.

2. Introduction

This year is a challenging year for the fashion industry. Because of the influence of coronavirus, people tend to shop more frequently online than in-store, so as a fashion trend analyst, which is my career goal, one needs to think about what lead consumers' to make shopping decision when they do the online shopping. Unlike shopping in stores, online shopping has its unique characteristics, people not only consider the product itself, product ratings and sales performance may also change one's mind of whether or not to buy this product.

Sales of summer clothes in E-commerce Wish dataset is from Kaggle.[^1] The dataset comes from Wish, which is a shopping website, the products listed in the dataset are products would appear if you search "summer" in the search area of the platform. The dataset summer products with rating and performance contain summer-related products available for sale, as of July 2020, this dataset contains 1573 observations with 43 variables.

This project aims to detect the attribution of the success of a product in online shopping, which factors have a strong effect on people to made decisions.

3. Method

3.1 Data Preparation

The summer products dataset contains 1573 products from Wish and 43 variables. Among the 43 variables, 22 are numerical, 19 are string and 5 are categorical variables. I first select some variables of interest from them:

- price: price you would pay to get the product in EUR

- retail_price: reference price in EUR for similar articles on the market, or in other stores/places. Used by the seller to indicate a regular value.
- units_sold: Number of units sold
- uses_ad_boosts: Whether the seller paid to boost his product within the platform
- rating: Mean ratings of the product
- rating_count: Total number of ratings of the product
- badge_product_quality: Whether the product has a product quality badge
- shipping_option_price: Price of shipping
- shipping_is_express: Whether shipping is express
- countries_shipped_to: Number of countries the product can be shipped to
- has_urgency_banner: Whether the product has an urgency banner
- merchant_rating_count: Total number of ratings for the merchant selling the product
- merchant_rating: Mean merchant rating
- merchant_has_profile_picture: Whether the merchant has a profile picture

There are 45 products that have no ratings. I simply exclude them from the data because the sample size is large (1573) and the proportion of missing value is small.
I believe that a discount on the retail price is an incentive for many customers. Therefore, I create a variable *discount* which is the price the customer would pay divided by the reference price and exclude the retail_price variable.

3.2 Model Selection

The dependent variable in my model is a count variable, so I first choose the poisson regression in the model selection, fit a poisson model using `stan_glm` function. Next, I check the model overdispersion by using the `dispersiontest` function in package `AER`. Based on the results, the model's dispersion is 4822, which is greater than one, indicating overdispersion.

I also ran negative binomial regression, linear regression and another linear regression model after taking log of the number of units sold by using `stan_glm` function from package `rstanarm`, to find which model is suitable. Also, visualize binned plot residuals to find which model fits better. The figure is listed below.

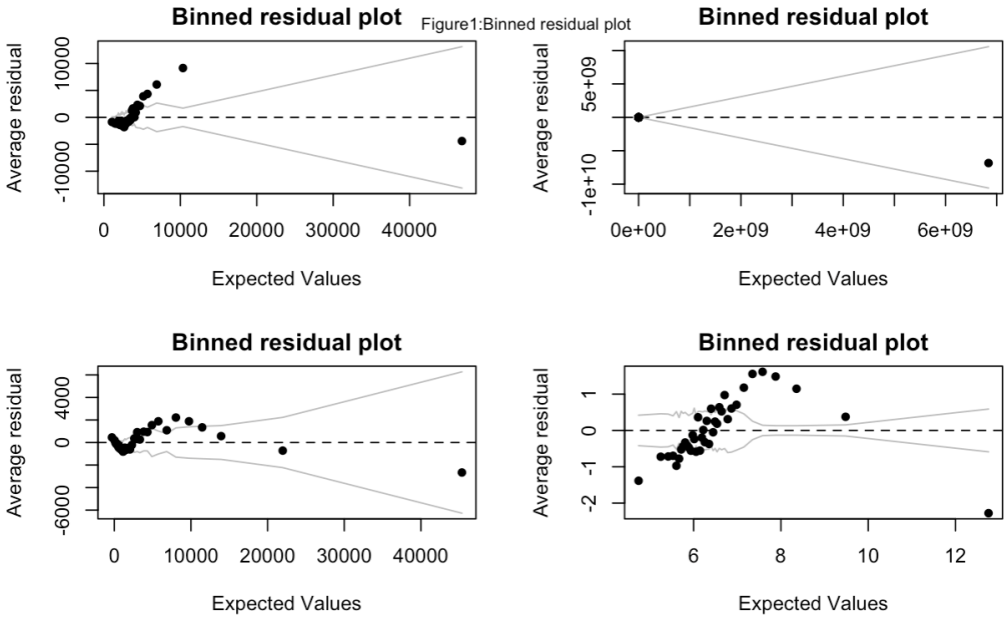


Figure1:Binned residual plot

Based on the binned residual plot, although all of the four models have points leave outside the confidence limits, the linear regression has smaller errors which are closer to the confidence limits, so I choose this model.

In this model, the outcome variable is the units_sold and predictor variables are price, discount, uses_ad_boosts, rating, rating count, badge_product_quality, shipping_option_price, shipping_is_express, countries_shipped_to, has_urgency_banner,merchant_rating_count,merchant_rating and merchant_has_profile_picture.

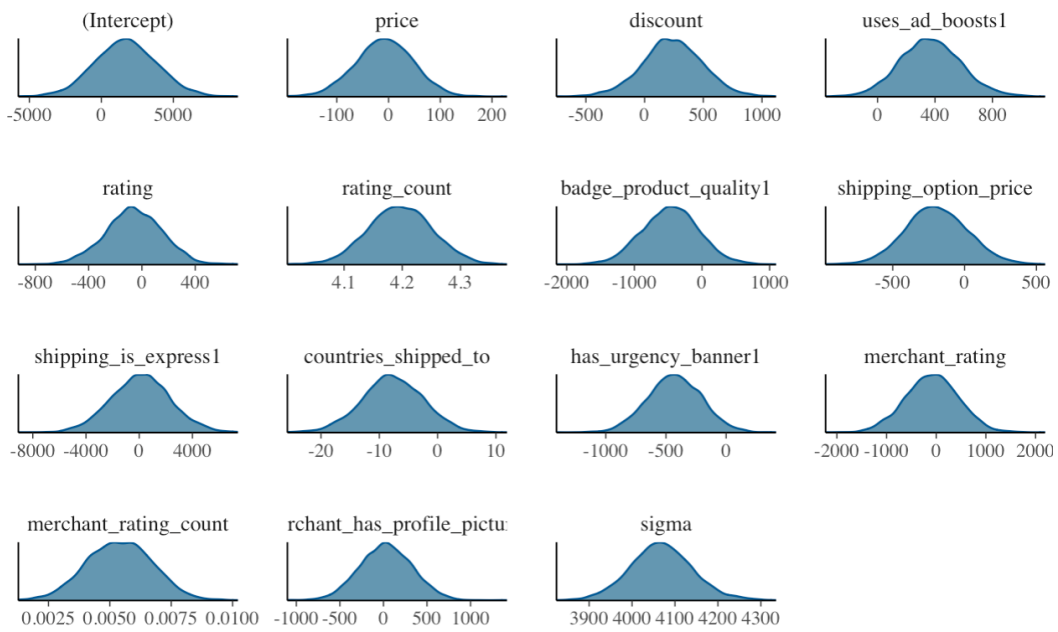
In the result part, I will discuss the results both in the model coefficient and estimates and model checking.

4. Results

4.1 Model Coefficients & Estimates

First, use the function `as.matrix()`, which allows us to access the matrix of posterior simulations to express uncertainty about the estimate of parameters.

Then, use the function `mcmc_dens()` in the `bayesplot` package, which allows looking at the distribution of the coefficients.



Also, I use `rope` function from `bayestestR` package to test the significance by checking the part of the credible interval that falls inside the ROPE interval. p in the below charts means get parameters from fit3, which is the model we choose in the previous part.

	Proportion of samples insi de ROPE [-0.10, 0.10]
rope(p$price)	0.11%
rope(p$discount)	0.00%
rope(p$uses_ad_boosts1)	0.00%
rope(p$rating)	0.00%
rope(p$rating_count)	0.00%
rope(p$badge_product_quality1)	0.00%
rope(p$shipping_option_price)	0.06%
rope(p$shipping_is_express1)	0.00%
rope(p$countries_shipped_to)	0.62%
rope(p$has_urgency_banner1)	0.00%
rope(p$merchant_rating)	0.03%
rope(p$merchant_rating_count)	100%
rope(p$merchant_has_profile_picture1)	0.03%

For variables discount, uses_ad_boosts1, rating, rating_count,badge_product_quality1, shipping_is_express1 and has_urgency_banner1, almost all the credible interval is outside the ROPE range, which means those coefficients are highly significant. For merchant_rating_count, almost all credible interval is inside the ROPE range, which means the coefficient is not significant.

4.2 Model Checking

Draw simulated predictions by using `posterior_predict()`function and using `the ppc_dens_overlay()` function to do the data visualization to see whether the data's distribution fall within the simulation. The model checking figure is in the appendix.

Although the model does capture some of the patterns but loses a lot of patterns as well. The model doesn't reach the peaking point of the data, and it misses a lot of 0's from figure 2 above. This means our model still needs some improvements, and I will discuss it in the discussion part.

5. Discussion

There are lots of things in my results that are consistent with real life. For example, discount, rating, quality of a product, shipping, and urgency banner of a product all are things we considered when we shopping online, and this analysis result shows that these factors have a strong correlation with an online product's success and affect one's behavior of buying something. And I also find that the total number of ratings for the merchant selling the product doesn't have a large effect on whether one buys products or not, because some ratings of a product are fake, and consumers seem to know that marketing method of merchants.

However, when I check my data, I find some confusion with my dataset. The unit_sold variable in my dataset, which is also the response variable I used in model fitting, is not the actual value, all the value of units sold in my dataset are integer and part of them are pretty large as well. Thus, because of the limitation of the units_sold value in my dataset, the model I fit and model checking occurs some error. Also, the dataset I use is from a shopping website called Wish, which is not a widely used website, so some results may not have universally.

Based on those limitations, my model still needs some improvements. For example, find a dataset from a widely-known website and be careful on the value in the dataset, because these values need to meet what happened in our actual life.

6. Bibliography

1. Kaggle: Sales of summer clothes in E-commerce Wish. https://www.kaggle.com/jmmvutu/summer-products-and-sales-in-ecommerce-wish

2. Goodrich, Ben; Gabry, Jonah; Ali, Iamd; Brilleman, Sam (2018). "rstanarm: Bayesian applied regression modeling via Stan." R package version 2.17.4, http://mc-stan.org/.

3. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

4. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data

Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

5. Gabry J, Mahr T (2020). "bayesplot: Plotting for Bayesian Models." R package version 1.7.2.

https://mc-stan.org/bayesplot.

6. Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). bayestestR: Describing Effects and their

Uncertainty, Existence and Significance within the Bayesian Framework. Journal of Open Source

Software, 4(40), 1541. doi:10.21105/joss.01541

7. Appendix

7.1 Exploratory Data Analysis

The question in this project is to find which factors have a strong effect on people to make decisions. The response variable is the number of units sold. A larger number of units sold means more people ended up buying this product. The number of units sold is not exact numbers: they were either rounded up or rounded down to 2, 7, 10, 50, 100, 1000, 5000, 10000, 20000, 50000 or 100000. Below is the relationship between the number of units sold and other variables.

Figure 2:relationship between price and number of units sold

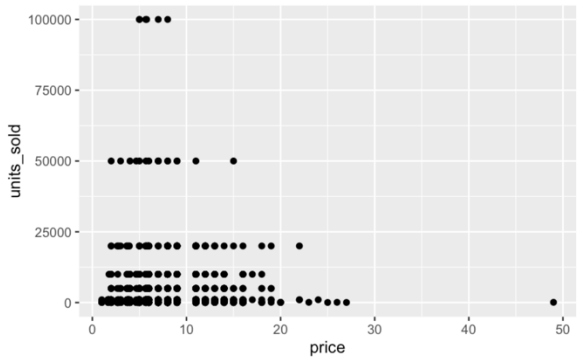


Figure 3:relationship between rating and number of units sold

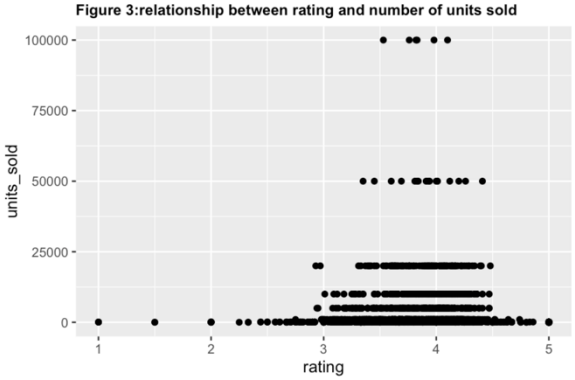


Figure 4:relationship between rating count and number of units sold

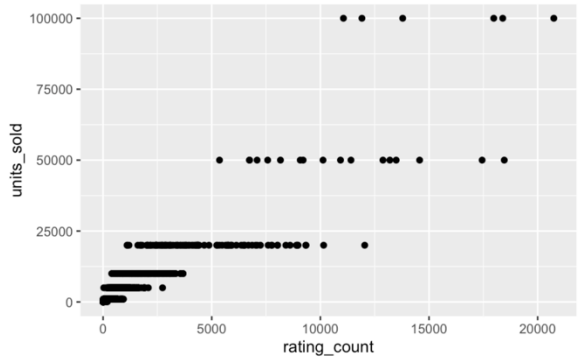


Figure 5:relationship between merchant rating and units sold

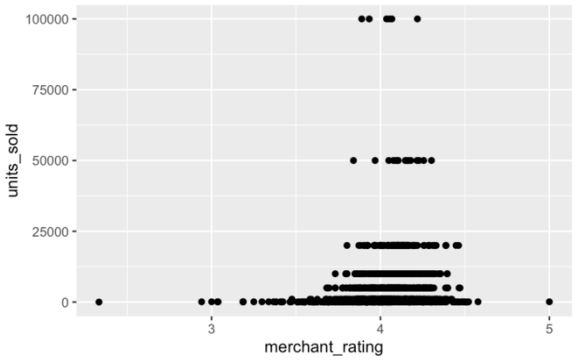
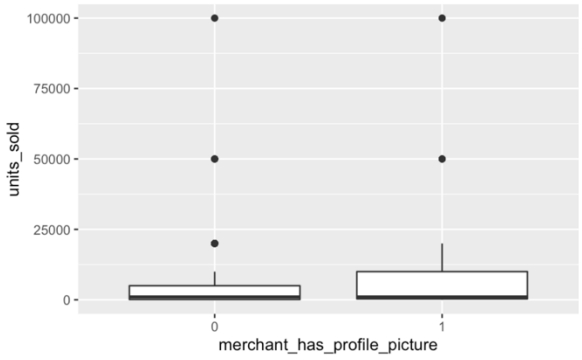


Figure 6:marchant has profile picture vs units sold

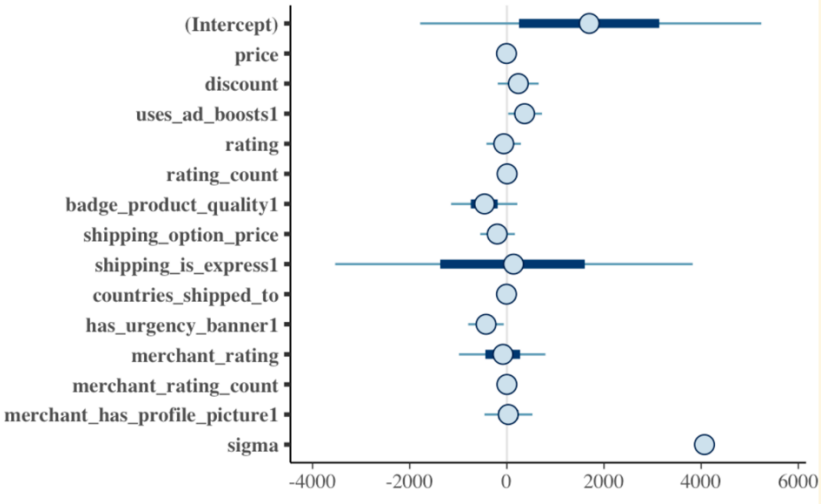


In the scatter plots of the number of units sold against other variables, we can find that units_sold is positively related to product rating, rating count and merchant rating. Units_sold is negatively related to price. The boxplot of units_sold by whether the merchant has a profile picture suggests that merchants with a profile picture may sell more products. A t-test comparing the two merchants with a profile picture to those who do not give a p-value less than 0.05, indicating that there is a significant difference in the number of units sold among them. This is an interesting finding suggesting that merchants should use a profile picture to sell more products.

7.2 Posterior intervals (credible intervals) visualization

visualize the posterior intervals (credible intervals) using `mcmc_intervals()`

Figure 7:Visulization of posterior intervals



7.3 Model Checking

Figure 8 : Model Checking