Predicting AsiaYo Users' Spending for Improved Search Results

Final Project Report

National Tsing Hua University

Business Analytics using Data Mining

Fall 2017

Submitted by:

Travis Greene, Martin Hsia, Letitia She, Leo Lee

Executive Summary

I. Business Problem- How can we improve our conversion rate?

AsiaYo earns revenues by taking a fixed percentage of total booking costs, so if we could increase the number of bookings and the amount spent per user, we would directly increase sales revenues. In order to achieve this end, we propose sorting user searches based on an estimated budget, with properties closest to the predicted nightly budget listed first in the most prominent area of the screen. Currently, the sorting algorithm returns search results with a very wide range of prices. We believe that a better sorting algorithm will contribute to a better user experience and ultimately to an improved conversion rate, especially under the assumption that AsiaYo customers make booking decisions largely based on price.

II. Data

In order to build a statistical learning model that can predict users' nightly spending budgets, we obtained a dataset of completed past transactions. Each row in the dataset consisted of a completed booking on AsiaYo. After cleaning, filtering, and manipulating the data into a usable form, there were approximately 50,000 rows of transactions and 16 variables (columns). Key variables used as input into the model included day of check in, day of booking order, user country, accommodation city and country, number of guests, and month of booking order. Our output column was the nightly spending amount.

Finally, after running several predictive models trained on the data, the most important variables associated with per night spending were the day of check-in (Saturday), accomodation city, particularly those in Japan, and the number of days booked in advance. We suspect this is because travelers to Japan and Korea spend more on average and also book longer in advance.

III. Analytics Solution

Our best performing model consisted of an ensemble model that used three separate models as input and gave us a predicted nightly spending budget as output. Though slightly more complicated to build, the ensemble's predictions were more accurate compared to any of the individual models. We believe this tradeoff in speed and complexity was necessary as the predictions are only useful if they relatively accurate. The time needed to compute a new user's predicted nightly budget is still negligible.

Using root mean squared error (RMSE) as our performance metric, the ensembled model scored 846. In more concrete terms, this means the difference between a customer's true nightly spending and our prediction was around $846, on average. Further, by creating a Taipei-level model, we found that RMSE could be reduced to nearly $600. We are confident that with careful tuning of the default parameters of the input models and the creation of city-level models (based on the top 1-3 cities by country), we could generate predictions accurate enough to have a tangible impact on the booking conversion rate and user experience.

IV. Recommendations

We expect that with more detailed information about accommodation cities, such as city district, we could produce even more accurate predictions. However, it is not clear how we could use this information at the time of search, given that the current AsiaYo page only allows for the selection of city, and not districts. Additionally, if we could connect a user on the search page with her previous booking history, we could improve our predictions immensely. This past transaction data could also be linked with textual information in the form of user reviews and ratings. Overall, we hold the opinion that user budget predictions could be an effective input into a broader search results algorithm. Given the relatively low cost, ease of implementation, and the potential upside in user conversion, we regard this as a worthwhile business project for AsiaYo.

Detailed Report

I. Problem description

We want to minimize the number of users who search, but then give up because they cannot quickly find a property that fits their budget (i.e., conversion friction). From the perspective of the business, our algorithm presents a chance to increase the conversion rate of users who search but don't ultimately book; and from the user's' perspective, it means a more pleasant booking experience because we will put the most appropriate properties first in their search results.

Previously, search results were determined by manual selection (by AsiaYo staff) and hosts' location in a host database. Our goal is to intelligently sort this database and display those properties first whose prices fit the users' predicted spending budget.

II. Data description

AsiaYo graciously provided us with around 60,000 rows of transaction data of past user purchase/booking history. After cleaning, we had approximately 50,546 rows and 16 columns of data (see screenshot in Appendix Figure 1). After cleaning and removing variables that would not be available to us at the time of prediction (when the user clicks "Search"), we were left with 15 predictor variables.

Every row in the data represents a transaction made by a user. Our output variable is the amount that each user transaction spent per night. For instances where customers booked more than one room, we simply divided the total amount spent by the number of nights in order to get an average per night spending amount. This derived variable became our target variable "per_night" that we then tried to predict using our statistical learning model. It is of interest to note that we do not have any particular user's personal data. All of our data is at an aggregated level, and so is, in a sense, anonymized. This is partly due to AsiaYo's policy of data minimization and collection and also partly due to the fact that variables such as gender and age were simply missing too many values to be useful.

Below are several examples of exploratory data analysis, which might indicate some interesting points for further research.

1. *How far do customers book in advance, on average?* (see Appendix Figure 2)

We can see that when people travel to Japan or Korea, they tend to book more than 2 months ahead. This may be due to many things. For example, it could be because higher room prices per night in these countries or popular travel sites that are crowded in a specific time of the year. It would be interesting to look at the booking leads times for important holidays, such as Chinese New Year or Tomb Sweeping Day.

2. *PC/Mobile booking by cities in Taiwan* (see Appendix Figure 3)

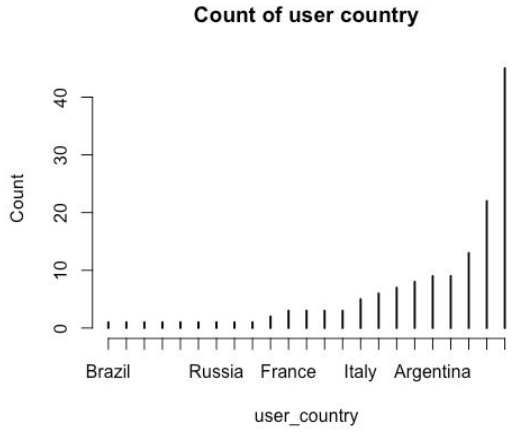
We discovered that the PC/Mobile booking rate has a consistent 80/20 split across Taiwan . This suggests that, for the most part, AsiaYo's mobile penetration is fairly evenly distributed across the island. However, it is interesting to note that Hsinchu and Penghu have zero mobile bookings. This is probably just because of relatively low number of bookings in those locations.

3. *Average days booked in advanced, cities in Taiwan* (see Appendix Figure 4)

Similar to booking in a foreign country, we can see that Matsu, which is a island near Taiwan, averages around 38 days of lead booking time. This makes sense because reaching Matsu is not so easy or convenient, and it's likely customers would need to plan transportation and lodging far in advance. AsiaYo may want to encourage cities at the low end (Hsinchu, Miaoli, Taichung) to book earlier, perhaps by offering early bird discounts. Earlier bookings would reduce the rate of host cancellations. This is because host cancellations often occur when a host has his property listed on multiple booking sites and forgets to de-list it when someone books on another platform. If AsiaYo can get its customers to book trips on AsiaYo before other platforms, then this will reduce the number of host cancellations.

III. Data Preparation

Step 1 : We started by first filtering for minimum counts of certain categorical values. For example, in the "user_country" column, we removed those values with less than 50 counts, such as Brazil, Russia, etc. We ran alternative models that specified minimum counts for transactions (i.e., only include cities if > 50 transactions there). Below you can see that there are many user countries with fewer than 50 counts. These "outliers" had to be removed because they caused some problems in our models' variable importance scores and during cross-validation.



Step 2 : Next, we turned all categorical variables into dummy variables, like accom_country into accom_country_Taiwan, accom_country_Japan, accom_country_Korea, etc. Therefore, we turned 16 columns to 146 columns. At this point we also converted all the variables with character values into factors, and all dates into date/time objects so that days, weeks, and times could be extracted and used as input. We ended up binning the create_at time into four six-hour bins in order to see if the time of day had any effect on the spending amount.

Step 3 : After that, we ran a recursive feature selection procedure in the R package "caret" that revealed that only about 130 out of the 150 variables reduced the model's RMSE (see Figure 5).

Step 4 : Finally, after converting all of our dates into date-time objects, we were able to derive another variable "days_book_before" that showed how many days in advance the customer booked the property.

IV. Data mining solution

Because we hypothesized that showing users overpredicted versus underpredicted properties might affect their booking behavior, we wanted to look at both RMSE and average error initially. The average error helped to see which models tended to over or underpredict. Ideally, we would choose a model that underpredicted budgets so that users would see properties below their actual budgets. We believed this to be better than showing users properties above their budgets. In the end, it was not a major issue as all of our models had prediction errors that were similar (see Appendix Figure 6 & 7).

The naive benchmark recorded an average error of -9.73, and the RMSE was 1347.47. We also tried three different packages in R for improved results. The results are shown below:

MODEL	GLMNet	GBM	Mars
Test Set RMSE	930.37	868.19	869.14

	Generalized Linear Model with Stepwise Feature Selection	Bagged MARS	Generalized Boosting Method
Coefficients	0.14514	0.03482	0.84898

Finally, we combined these results into an ensemble model with the coefficients above, recording an RMSE of 846.08, a 36% improvement compared to the naive RMSE. A scatterplot of the final ensemble's predictions against the observed values in the test set can be found in the appendix (see Figure 8). It is interesting to note that our model tends to overpredict at the low end of the per night variable, but then more evenly over and underpredicts at higher spending levels.

V. Conclusion and Recommendations

Overall, we believe AsiaYo can improve their default sorting algorithm by matching users with a property that closely matches their predicted budget. Further, by combining several metrics such as spending per night, rejection rate of hosts, and others, AsiaYo can achieve better predictions and thus better sorting results. Nevertheless, AsiaYo's current policy minimizes customer data collection in order to respect users' privacy concerns. The result is a trade-off between accuracy of prediction and maintaining users' trust.

Regarding future operational recommendations, we would like to see more detailed user data, especially demographic data. We also believe that the UTM_Source variable (the origin of the user, whether from Facebook, or Google, or blogs, etc.) could be a useful predictor of per night spending. Unfortunately, most of the values were missing.

One solution to the problem of reducing RMSE would be to use city or country-level models. But one issue with this is that we currently have too few data for some cities and countries. In any case, we tested a Taipei-only model, which gave us an RMSE of approximately $600 NTD and a much narrower distribution of prediction errors (see Figure 9). This approach looks promising.

Given the difficulty of predicting a numeric outcome with sufficient accuracy, another potential approach would be to transform the regression problem into a classification problem that predicted user budgets as falling into either high, medium, or low buckets.

Finally, we would like to thank AsiaYo for generously providing us with the data and for supporting us throughout the project. We hope our results can help with the implementation of the new and improved AYSort.

Appendix

Figure 1

	guests	nights	rooms	platform	user_country	accom_country	accom_city	days_book_before
1	1	1	1	pc	Taiwan	Taiwan	Yilan_County	7
2	4	1	1	pc	Taiwan	Taiwan	Yilan_County	11
3	2	1	1	pc	Malaysia	Taiwan	New_Taipei_City	26
4	4	1	1	pc	Taiwan	Taiwan	New_Taipei_City	17
5	2	1	1	pc	Taiwan	Taiwan	Taipei_City	4

per_night	day_order	month_order	time_of_day	day_checkin	month_checkin	day_checkout	month_checkout
3800	3	9	3	4	9	5	9
3200	4	9	3	2	9	3	9
1380	4	9	3	3	10	4	10
2680	6	9	4	3	9	4	10
1380	3	9	2	7	9	1	9

(5 rows spit in half)

Figure 2

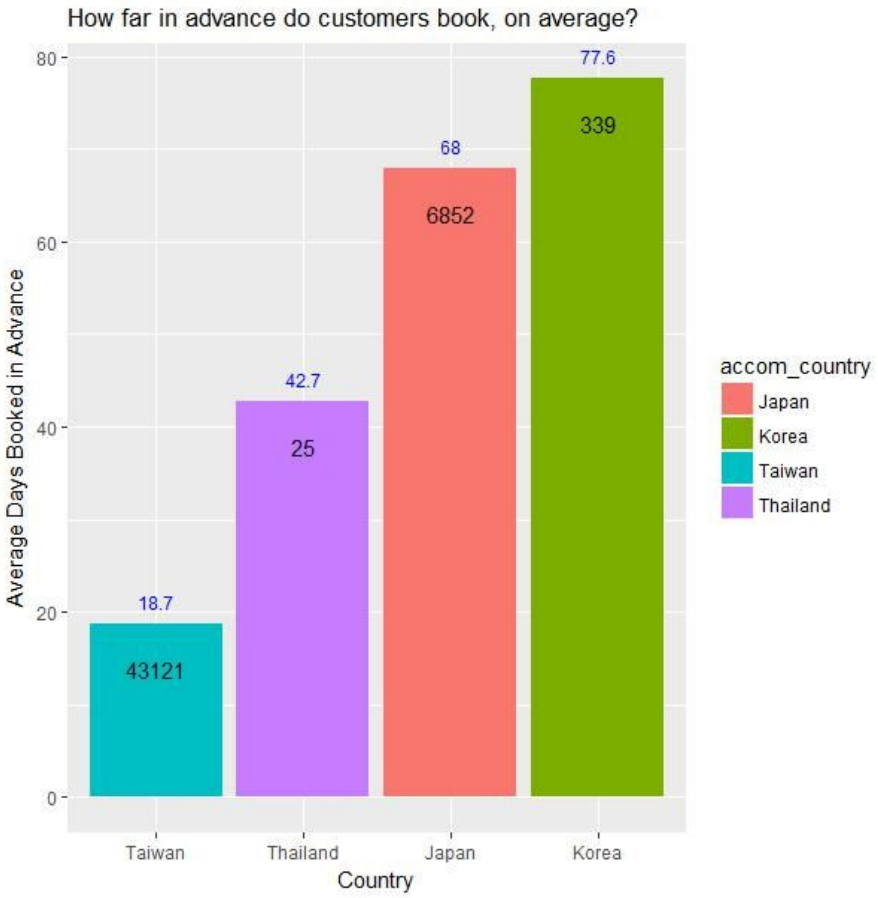


Figure 3

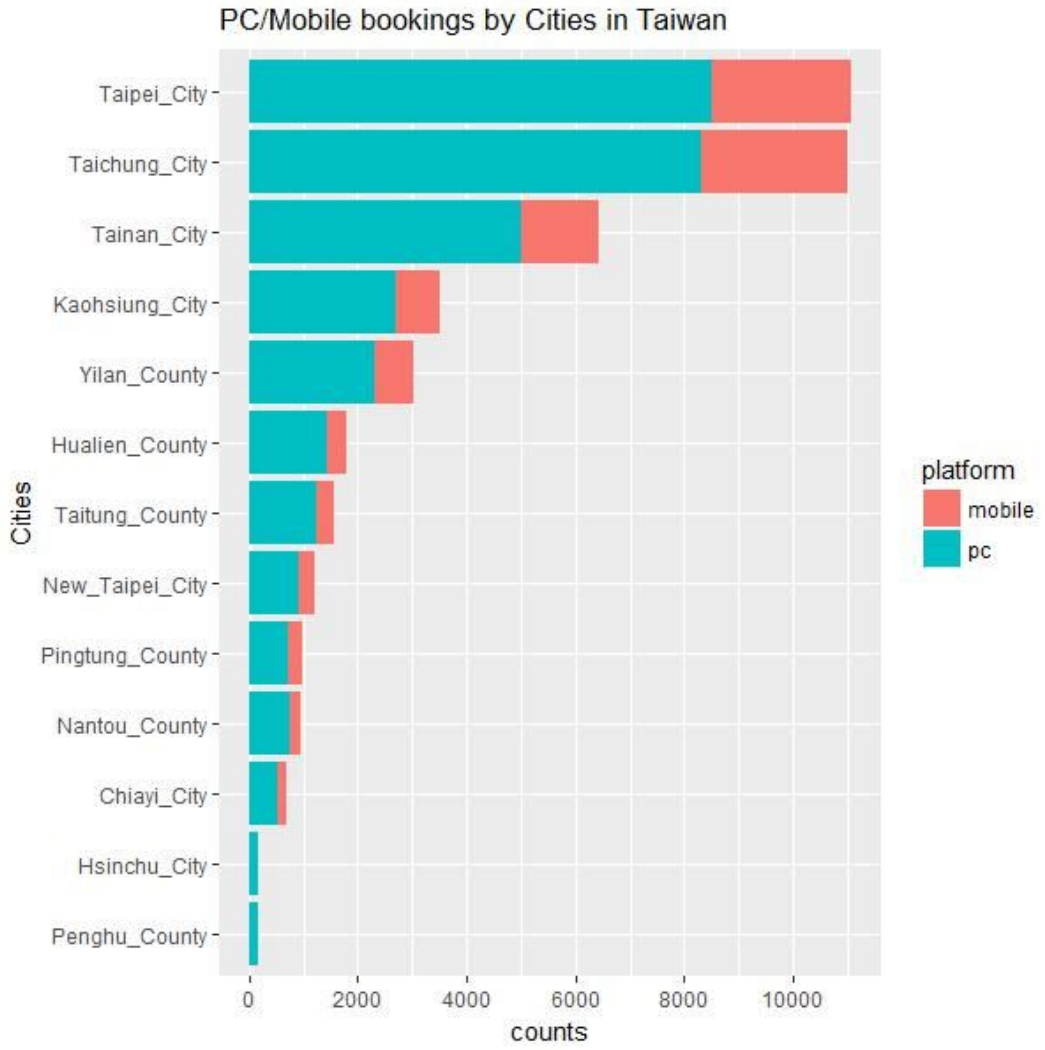


Figure 4

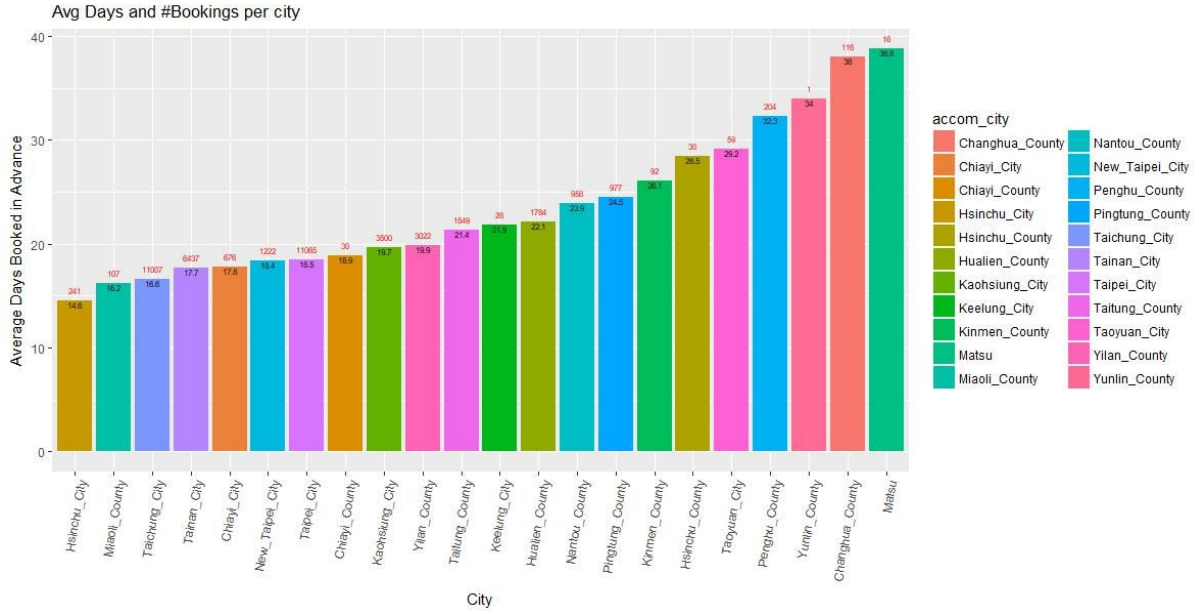


Figure 5

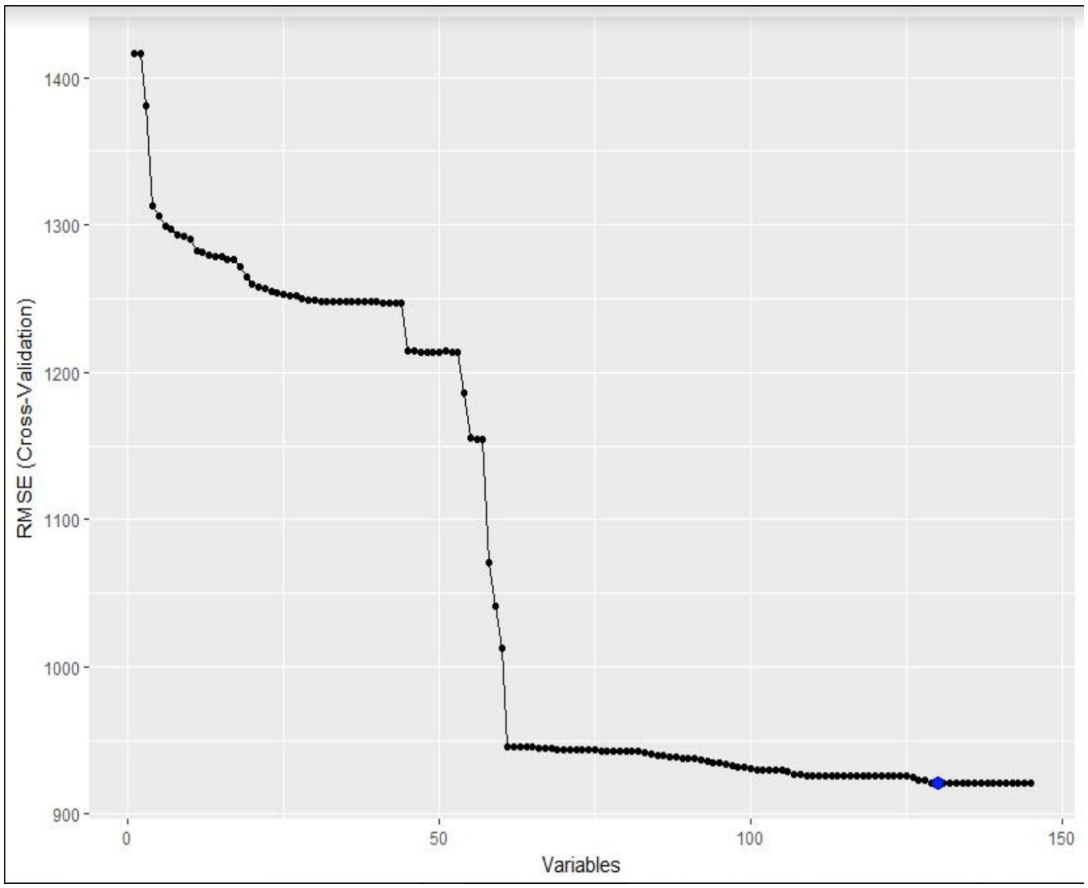


Figure 6

MODEL	Naive	Random Forest	Linear Reg.	Linear Reg. (best subsets & standardized)	Lasso Reg.	Ridge Reg.	K-NN Reg.	Regression Tree	Neural Net	Boosted Tree	Bagged Tree
AVG. error	-9.730594	-10.09135	-16.028	Froze :-(	-15.69396	-9334.517	-48.40446	-10.68459	-16.69444	24.4	18.92
Test Set RMSE	1317.471	877.0068	938.7423		948.274	11814.84	949.6159	916.5837	1100.889	884.43	908.95

Figure 7

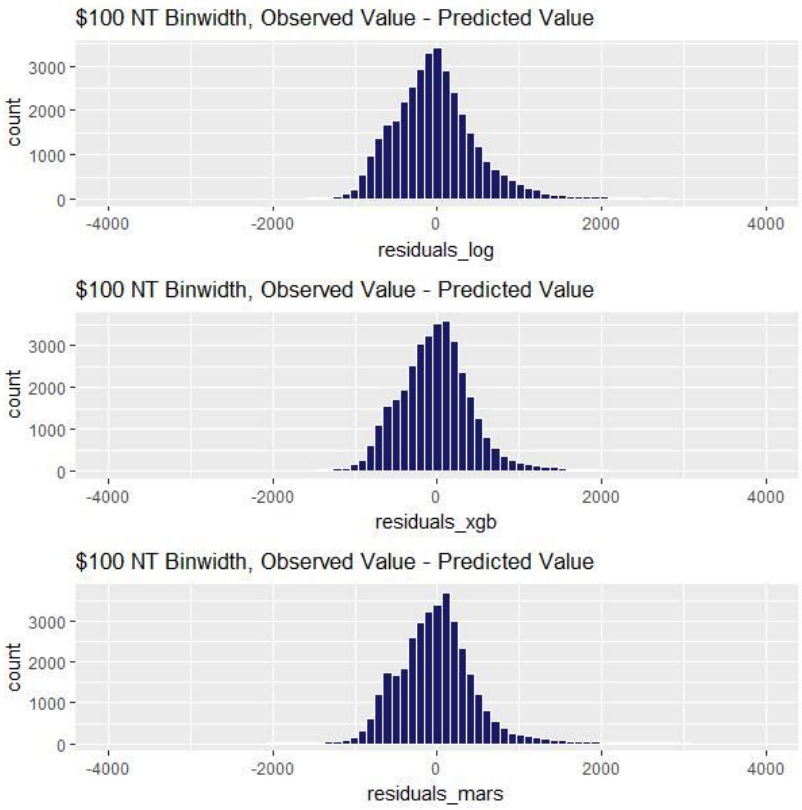


Figure 8

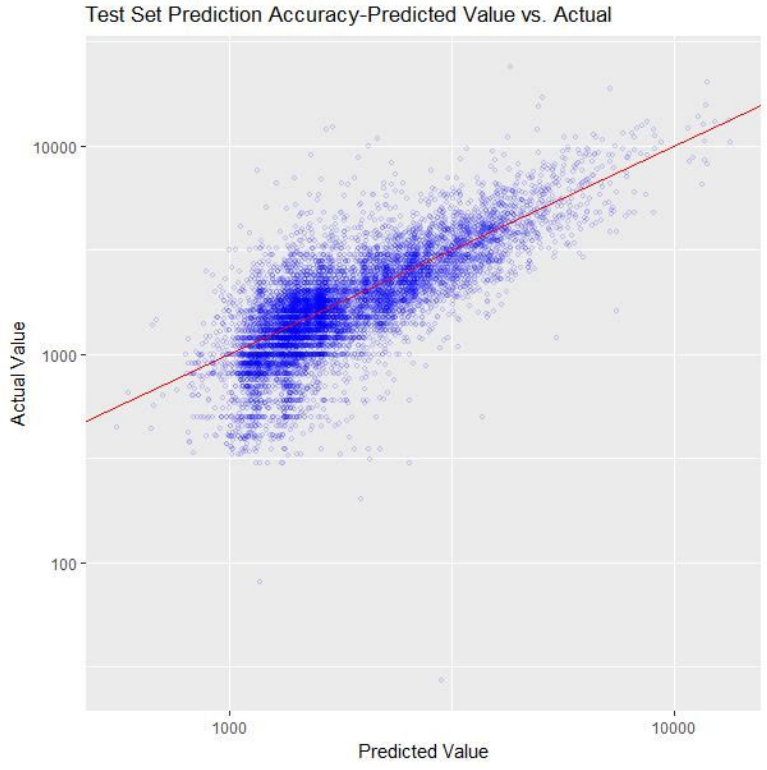
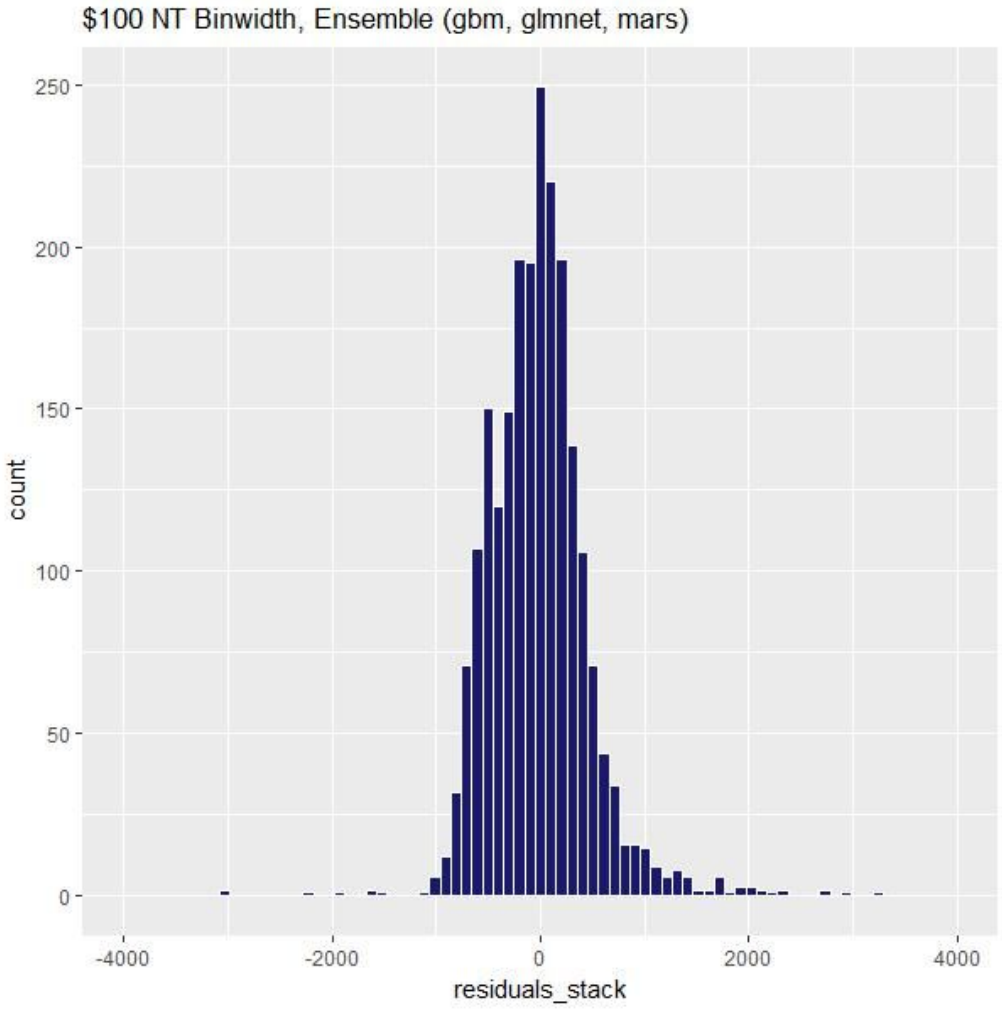


Figure 9


$100 NT Binwidth, Ensemble (gbm, glmnet, mars)
count
250
200
150
100
50
0
-4000
-2000
0
2000
4000
residuals_stack