ECO 395 Homework 2

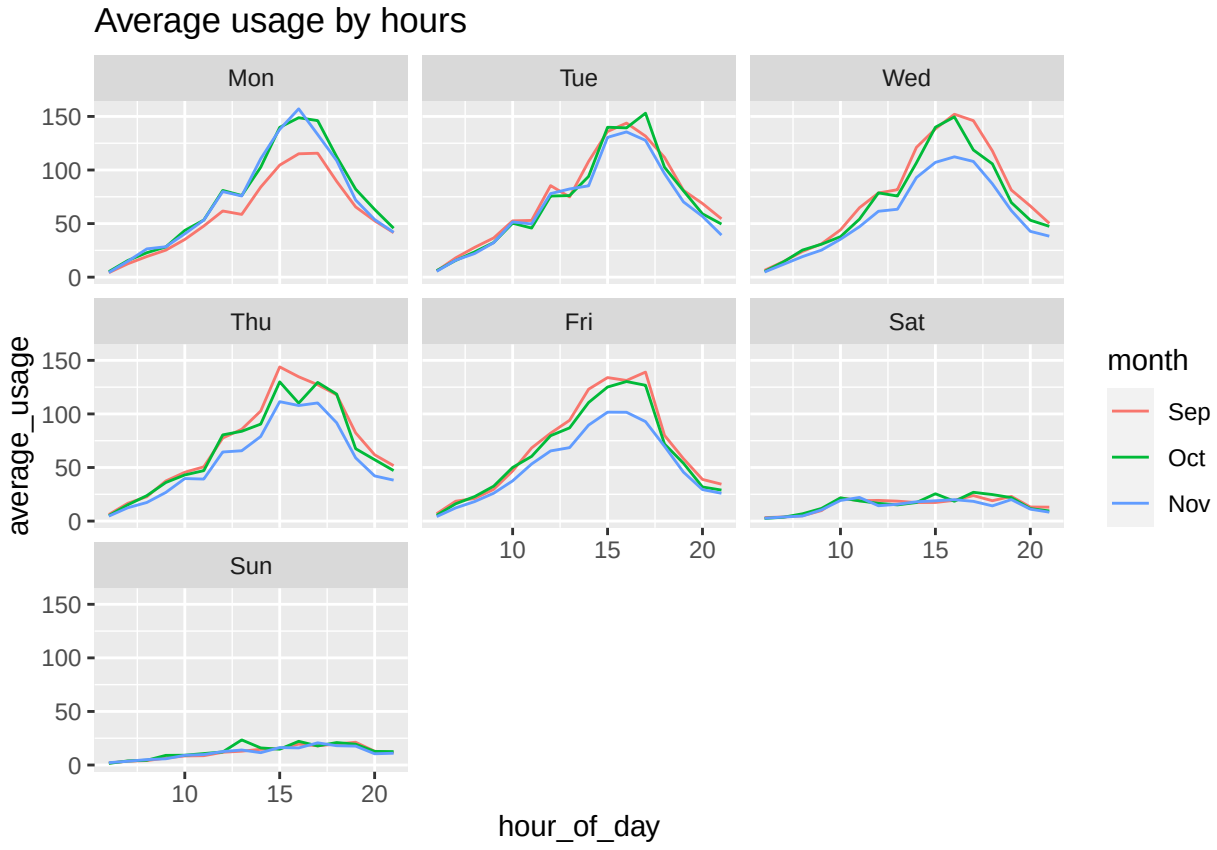
Yefu Chen

3/11/2021

QUESTION 1

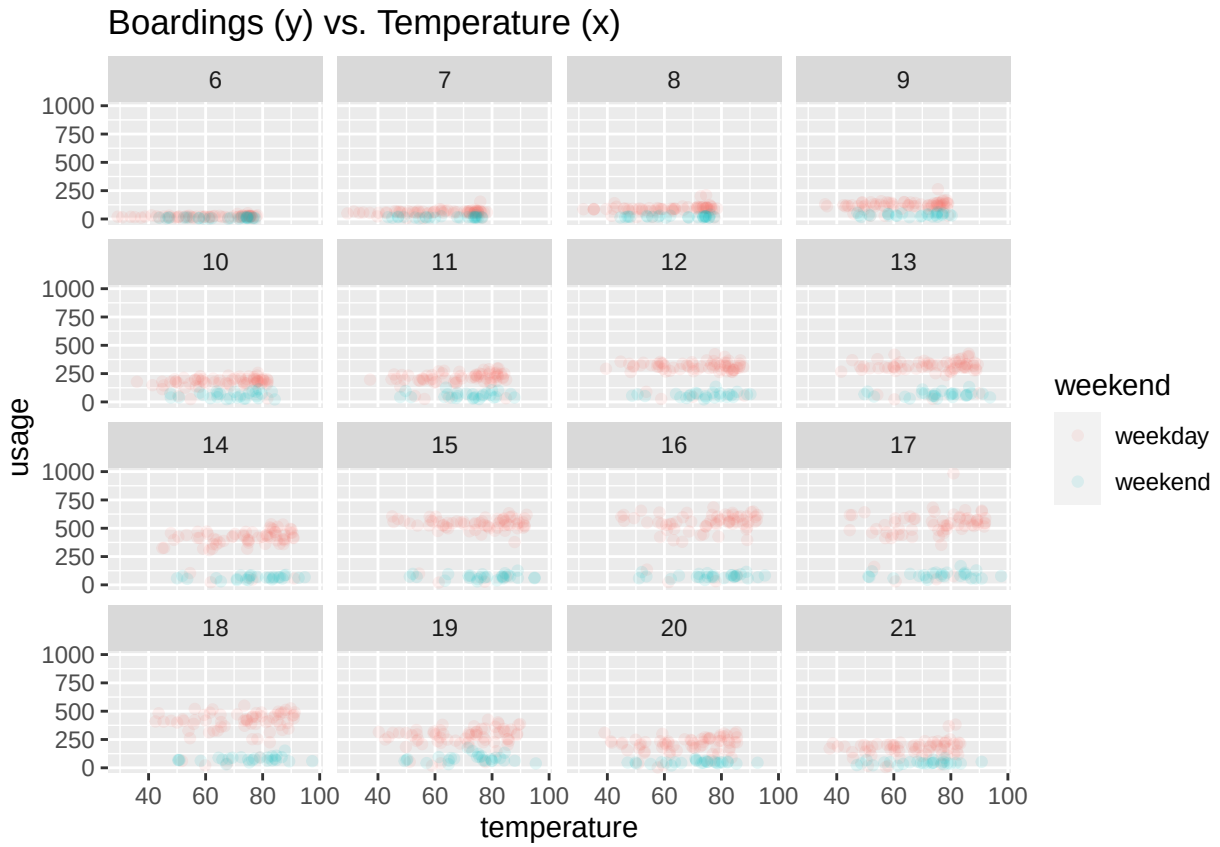
Your task in this problem is to make two faceted plots and to answer questions about them.

A) One panel of line graphs that plots average boardings grouped by hour of the day, day of week, and month.



Caption: This figure presents the faceted plot for Question A. The hours of peak boardings do not change from days to days. It is similar among weekdays, from Monday to Friday. But the average usage changes a lot from weekdays to weekends. I also notice that average boardings on Mondays in September look lower, and average boardings on Weds/Thurs/Fri in November look lower. I assume 1) in September, students usually stay on campus to have reunions and "shop" courses. Hence there may not have a lot of public transit demand. 2) in November, students may prefer to stay on campus and prepare for the midterm/final exam.

B) One panel of scatter plots showing boardings (y) vs. temperature (x) in each 15-minute window, faceted by hour of the day, and with points colored in according to whether it is a weekday or weekend.



Caption: This figure is a sort of scatter plots showing boardings (y) vs. temperature (x) in each 15-minute window, faceted by hour of the day, and with points colored in according to whether it is a weekday (red) or weekend (green). Temperature seems not to have a noticeable effect on the number of UT students riding the bus.

QUESTION 2

For this data set, you'll run a "horse race" (i.e. a model comparison exercise) between two model classes: linear models and KNN.

Answers: After testing, there is a good linear model (RMSE= 61715.89). The independent variables of this model include waterfront (Y/N), new construction status (Y/N), central air status (Y/N), heating types, years of building, lot size, number of rooms, number of bathrooms, number of bedrooms, and (livingArea + landValue)ˆ2. The positive indicators include living area, land value, number of bathrooms, new construction status, lot size, and number of rooms. The negative indicators include waterfront status, central air status, years of building, number of bedrooms, and heating types (stem and electric to hot air). Also, after testing, there is a good KNN regression (RMSE= 63676). The number of K is 7. In this study, the linear model performs better than the KNN regression model. Also, the linear model is easier to interpret and understand to the public.

QUESTION 3

What do you notice about the history variable vis-a-vis predicting defaults? What do you think is going on here? In light of what you see here, do you think this data set is appropriate for building a predictive model of defaults if the purpose of the model is to screen prospective borrowers to classify them into "high" versus "low" probability of default? Why or why not—and if not, would you recommend any changes to the bank's sampling scheme?

Answers: Figure 3-1 demonstrates the credit history vs. rates of fell into default. I notice that the rates of fell into default are higher in persons with good credit history than those with poor and terrible credit history. The results of logistic regression are consistent with this finding. The coefficient of credit history plays a negative role in predicting default status, which means that worse credit history can have lower rates of fell into default. Compared to good credit history persons, the odds of poor credit history felling into default decrease by 0.66, and the odds of a terrible credit history person decrease by 0.84. This result is not consistent with common sense that persons with good credit history tend to avoid default. One reason could be that when it is the first time to apply for a loan, individuals have a good credit history, while others may not repay the loan. Although this model has controlled the effects of age and purposes, it does not consider whether it is the first time apply for a loan. Hence, I do not think this dataset is appropriate for building a predictive model of defaults. I suggest that the bank consider the effects of first-time application to loan and separate the sample as a first-time group and multiple-times group. Also, the two hurdle model may be helpful in this prediction.

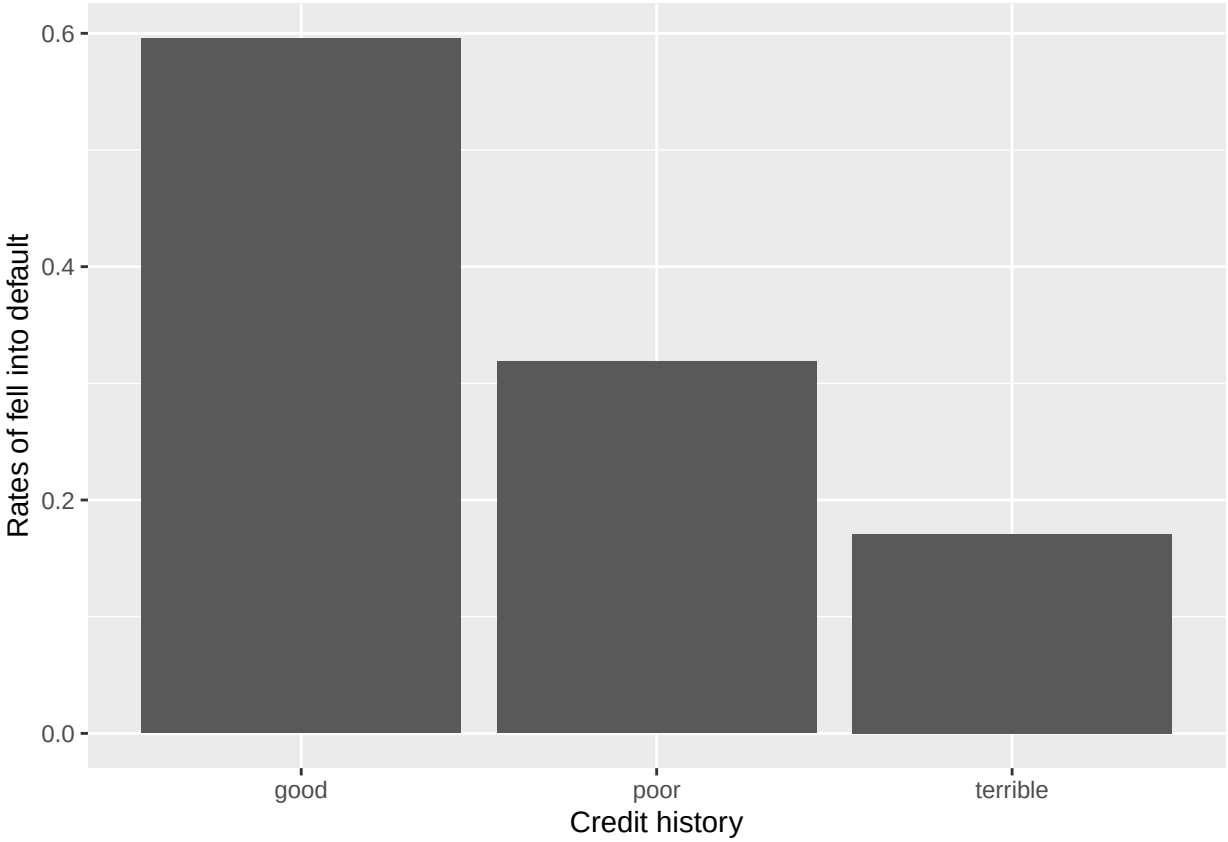


Figure 3-1. Credit history vs. Rates of fell into default

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.7075	0.4726	-1.497	0.1343
historypoor	-1.108	0.2473	-4.479	7.511e-06

	Estimate	Std. Error	z value	Pr(>|z|)
historyterrible	-1.885	0.2822	-6.679	2.407e-11
duration	0.02526	0.0081	3.118	0.001818
amount	9.596e-05	3.65e-05	2.629	0.008556
installment	0.2216	0.07626	2.906	0.00366
age	-0.02018	0.007224	-2.794	0.005208
purposeedu	0.7248	0.3707	1.955	0.05058
purposegoods/repair	0.1049	0.2573	0.4077	0.6835
purposenewcar	0.8545	0.2773	3.081	0.002063
purposeusedcar	-0.7959	0.3598	-2.212	0.02694
foreigngerman	-1.265	0.5773	-2.191	0.02849

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	1222 on 999 degrees of freedom
Residual deviance:	1070 on 988 degrees of freedom

QUESTION 4

Model building: Using only the data in hotels.dev.csv, please compare the out-of-sample performance of the following models. **Step 1**: Once you've built your best model and assessed its out-of-sample performance using hotels_dev, now turn to the data in hotels_val. **Step 2**: Next, create 20 folds of hotels_val and make predictions.

Since the RMSE is not applicable for binominal logisti regression, I use the accuracy as the index to quantify performance of models. My model is given by (children ~ reserved_room_type:meal + average_daily_rate + poly(total_of_special_requests,2) + assigned_room_type + market_segment + hotel + poly(adults,3) + booking_changes + customer_type + previous_bookings_not_canceled + poly(lead_time,3) + distribution_channel + is_repeated_guest + poly(stays_in_weekend_nights,3) +required_car_parking_spaces +poly(days_in_waiting_list,3)). The accuracy is 0.864, better than baseline 1 (0.661) and 2 (0.847).

Figure 4-1 present model validation step 1. It is a plot TPR(t) vs. FPR(t) as the classification threshold t=0.1, 0.5, 0.8.

Figure 4-2 is a barplot about the actual number of children vs. the predicted number of children across 28 folders. We can see in folders 15, 16, and 19, the model performance relatively good.

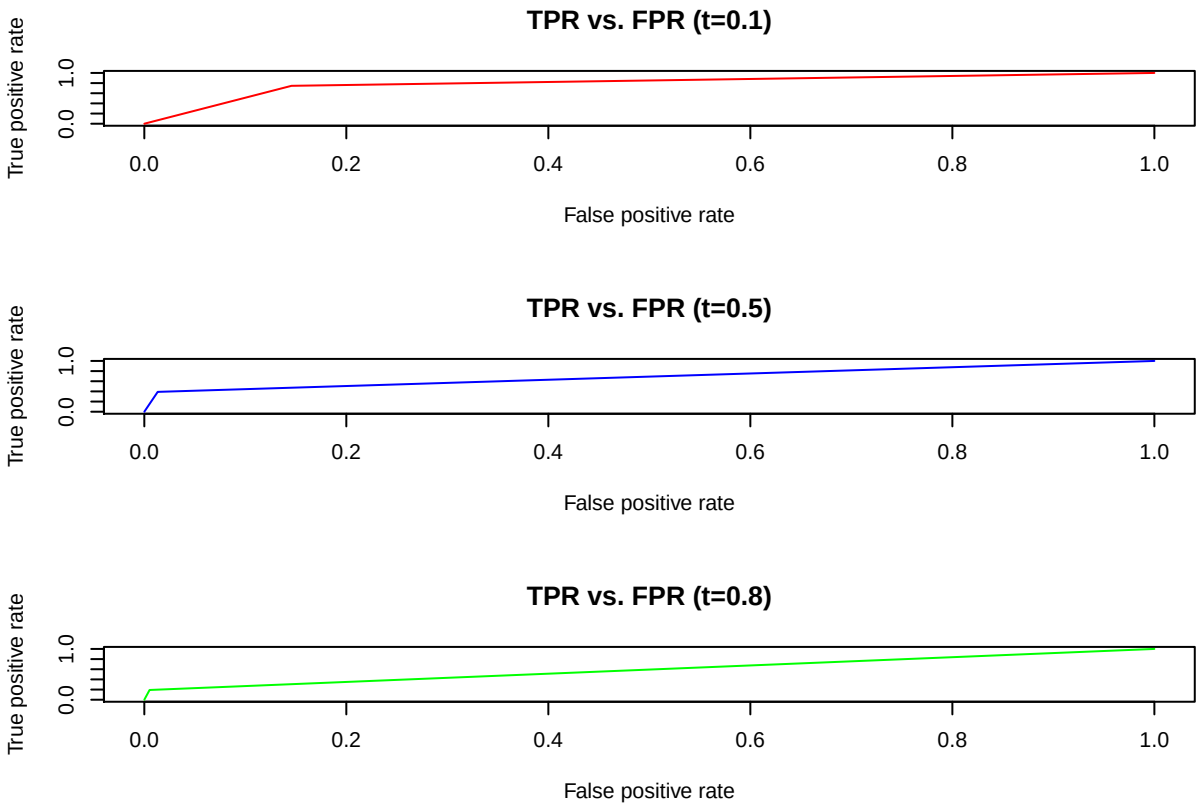


Figure 4-1. TPR(t) vs. FPR(t) when t= 0.1, 0.5, and 0.8.

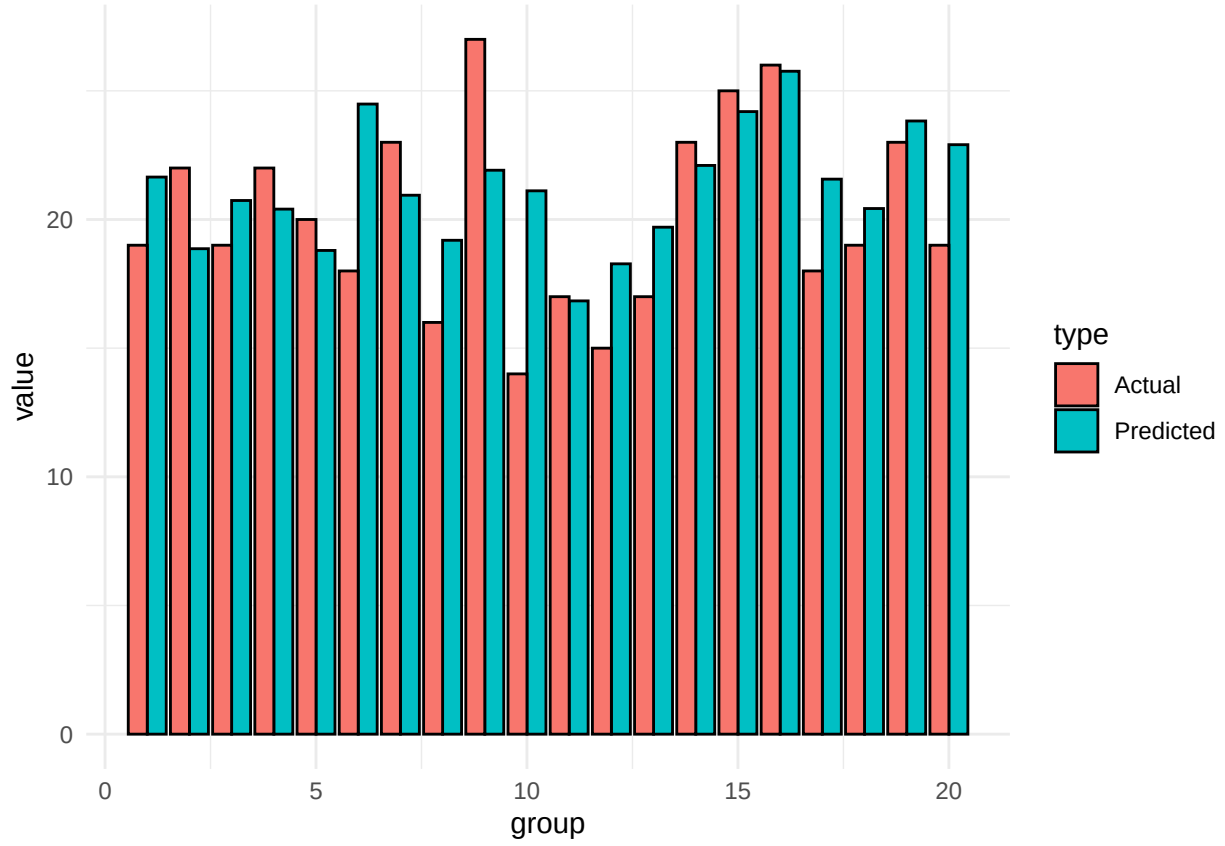


Figure 4-2. Predicted number (Green) vs. Actual number (Red)

Appendix

For QUESTION 2

I separately test the performance of the linear model and KNN regression. The RMSEE is the index to quantify performances. In the linear models, I first covert six categorical variables (heating, fuel, sewer, waterfrnont, newConstruction, and centralAir) into factor variables. Then, I modify the portion of test and train data, and find that 0.7 is the best. I apply two linear models. The first linear model (price ~ lotSize + bedrooms + bathrooms) performances fine (RMSE= 80153.51), and the second linear model (full models) performance better (RMSE= 62303.69). Afterwards, I apply Stepwise to find out the best combination as the third linear model. The result presents, price that when independent variables include livingArea, landValue, bathrooms, waterfront, newConstruction, heating, lotSize, centralAir, age, bedrooms, and rooms, the model performance better (RMSE= 61870.59). Then, I modify the model of Stepwise, introducing the **(livingArea + landValue)^2**, the model performance better (RMSE= 61715.89). Besides, I also tried log-transformations, polynomial terms, and different interactions, but this the last model (price ~ (livingArea + landValue)^2 + waterfront + bathrooms + newConstruction + centralAir + age + lotSize + rooms + heating + bedrooms) is the best.

For QUESTION 4

Table 3: Table continues below

	Estimate	Std. Error	z value
(Intercept)	-11.62	233.9	-0.04967
average_daily_rate	0.01143	0.0005148	22.21
poly(total_of_special_requests, 2)1	84.08	4.753	17.69
poly(total_of_special_requests, 2)2	-8.003	3.667	-2.183
assigned_room_typeB	0.3844	0.1763	2.18
assigned_room_typeC	1.7	0.1471	11.55
assigned_room_typeD	1.339	0.075	17.85
assigned_room_typeE	1.197	0.139	8.609
assigned_room_typeF	1.221	0.1722	7.091
assigned_room_typeG	1.408	0.2108	6.679
assigned_room_typeH	1.81	0.3763	4.811
assigned_room_typeI	1.665	0.3575	4.656
assigned_room_typeK	-0.7752	0.4561	-1.7
market_segmentComplementary	12.78	233.8	0.05464
market_segmentCorporate	11.68	233.8	0.04996
market_segmentDirect	12.28	233.8	0.05253
market_segmentGroups	11.38	233.8	0.04867
market_segmentOffline_TA/TO	12.81	233.8	0.05478
market_segmentOnline_TA	12.82	233.8	0.05482
hotelResort_Hotel	-0.995	0.06084	-16.35
poly(adults, 3)1	-27.59	6.885	-4.007
poly(adults, 3)2	-97.54	5.872	-16.61
poly(adults, 3)3	-92.61	4.852	-19.09
booking_changes	0.2903	0.02537	11.44
customer_typeGroup	-0.1839	0.3998	-0.4601
customer_typeTransient	0.4769	0.1387	3.437
customer_typeTransient-Party	-0.1647	0.1559	-1.056
previous_bookings_not_canceled	-0.3384	0.1161	-2.914
poly(lead_time, 3)1	-8.753	8.263	-1.059
poly(lead_time, 3)2	-42.37	10.4	-4.073
poly(lead_time, 3)3	-30.67	8.763	-3.499
distribution_channelDirect	0.3489	0.3581	0.9744
distribution_channelGDS	-13.21	237.3	-0.05569
distribution_channelTA/TO	-0.3305	0.3227	-1.024
is_repeated_guest1	-0.7112	0.2327	-3.057
poly(stays_in_weekend_nights, 3)1	-68.4	32.02	-2.136
poly(stays_in_weekend_nights, 3)2	-331.1	130.5	-2.538
poly(stays_in_weekend_nights, 3)3	-157.2	67.63	-2.325
required_car_parking_spacesparking	0.02036	0.07006	0.2906
poly(days_in_waiting_list, 3)1	-120.2	86.94	-1.383
poly(days_in_waiting_list, 3)2	-129.8	113.1	-1.147
poly(days_in_waiting_list, 3)3	-120.1	72.39	-1.659
reserved_room_typeA:mealBB	-5.691	1.546	-3.683
reserved_room_typeB:mealBB	-4.687	1.558	-3.008
reserved_room_typeC:mealBB	-2.569	1.559	-1.648
reserved_room_typeD:mealBB	-7.037	1.548	-4.547
reserved_room_typeE:mealBB	-6.188	1.548	-3.997
reserved_room_typeF:mealBB	-4.27	1.547	-2.759
reserved_room_typeG:mealBB	-3.457	1.535	-2.253
reserved_room_typeH:mealBB	-1.413	1.596	-0.885
reserved_room_typeL:mealBB	-19.59	1532	-0.01279

	Estimate	Std. Error	z value
reserved_room_typeA:mealFB	-4.009	1.578	-2.54
reserved_room_typeC:mealFB	-4.398	1.708	-2.575
reserved_room_typeD:mealFB	-5.239	1.595	-3.286
reserved_room_typeE:mealFB	-20.33	773.6	-0.02627
reserved_room_typeF:mealFB	-20.79	931.4	-0.02232
reserved_room_typeG:mealFB	-19.04	1359	-0.01402
reserved_room_typeH:mealFB	-5.556	2400	-0.002316
reserved_room_typeA:mealHB	-5.599	1.547	-3.619
reserved_room_typeB:mealHB	-5.432	1.884	-2.884
reserved_room_typeC:mealHB	-2.831	1.587	-1.784
reserved_room_typeD:mealHB	-6.725	1.552	-4.333
reserved_room_typeE:mealHB	-6.369	1.555	-4.095
reserved_room_typeF:mealHB	-5.51	1.562	-3.527
reserved_room_typeG:mealHB	-2.933	1.554	-1.888
reserved_room_typeH:mealHB	-1.487	1.672	-0.8894
reserved_room_typeA:mealSC	-7.053	1.551	-4.548
reserved_room_typeB:mealSC	-5.493	1.896	-2.897
reserved_room_typeC:mealSC	-2.618	1.986	-1.318
reserved_room_typeD:mealSC	-6.84	1.674	-4.087
reserved_room_typeE:mealSC	-21.18	518.3	-0.04087
reserved_room_typeF:mealSC	-3.285	1.721	-1.909
reserved_room_typeG:mealSC	-18.82	1530	-0.0123
reserved_room_typeA:mealUndefined	-5.041	1.59	-3.171
reserved_room_typeC:mealUndefined	-4.185	1.707	-2.452
reserved_room_typeD:mealUndefined	-5.117	1.63	-3.139
reserved_room_typeE:mealUndefined	-5.431	1.893	-2.868
reserved_room_typeF:mealUndefined	-21.94	749	-0.02929

	Pr(>|z|)
(Intercept)	0.9604
average_daily_rate	2.643e-109
poly(total_of_special_requests, 2)1	5.06e-70
poly(total_of_special_requests, 2)2	0.02906
assigned_room_typeB	0.02925
assigned_room_typeC	7.129e-31
assigned_room_typeD	2.785e-71
assigned_room_typeE	7.386e-18
assigned_room_typeF	1.336e-12
assigned_room_typeG	2.413e-11
assigned_room_typeH	1.502e-06
assigned_room_typeI	3.219e-06
assigned_room_typeK	0.0892
market_segmentComplementary	0.9564
market_segmentCorporate	0.9602
market_segmentDirect	0.9581
market_segmentGroups	0.9612
market_segmentOffline_TA/TO	0.9563
market_segmentOnline_TA	0.9563
hotelResort_Hotel	4.014e-60
poly(adults, 3)1	6.155e-05
poly(adults, 3)2	5.91e-62

	Pr(>|z|)
poly(adults, 3)3	3.309e-81
booking_changes	2.601e-30
customer_typeGroup	0.6454
customer_typeTransient	0.0005875
customer_typeTransient-Party	0.2908
previous_bookings_not_canceled	0.003572
poly(lead_time, 3)1	0.2894
poly(lead_time, 3)2	4.643e-05
poly(lead_time, 3)3	0.0004665
distribution_channelDirect	0.3299
distribution_channelGDS	0.9556
distribution_channelTA/TO	0.3057
is_repeated_guest1	0.002237
poly(stays_in_weekend_nights, 3)1	0.03268
poly(stays_in_weekend_nights, 3)2	0.01116
poly(stays_in_weekend_nights, 3)3	0.02007
required_car_parking_spacesparking	0.7713
poly(days_in_waiting_list, 3)1	0.1666
poly(days_in_waiting_list, 3)2	0.2512
poly(days_in_waiting_list, 3)3	0.09713
reserved_room_typeA:mealBB	0.0002309
reserved_room_typeB:mealBB	0.002631
reserved_room_typeC:mealBB	0.09938
reserved_room_typeD:mealBB	5.443e-06
reserved_room_typeE:mealBB	6.415e-05
reserved_room_typeF:mealBB	0.00579
reserved_room_typeG:mealBB	0.02427
reserved_room_typeH:mealBB	0.3762
reserved_room_typeL:mealBB	0.9898
reserved_room_typeA:mealFB	0.01108
reserved_room_typeC:mealFB	0.01002
reserved_room_typeD:mealFB	0.001017
reserved_room_typeE:mealFB	0.979
reserved_room_typeF:mealFB	0.9822
reserved_room_typeG:mealFB	0.9888
reserved_room_typeH:mealFB	0.9982
reserved_room_typeA:mealHB	0.0002954
reserved_room_typeB:mealHB	0.003932
reserved_room_typeC:mealHB	0.0744
reserved_room_typeD:mealHB	1.469e-05
reserved_room_typeE:mealHB	4.229e-05
reserved_room_typeF:mealHB	0.0004206
reserved_room_typeG:mealHB	0.05905
reserved_room_typeH:mealHB	0.3738
reserved_room_typeA:mealSC	5.421e-06
reserved_room_typeB:mealSC	0.003764
reserved_room_typeC:mealSC	0.1873
reserved_room_typeD:mealSC	4.377e-05
reserved_room_typeE:mealSC	0.9674
reserved_room_typeF:mealSC	0.05623
reserved_room_typeG:mealSC	0.9902
reserved_room_typeA:mealUndefined	0.001521

	Pr(>|z|)
reserved_room_typeC:mealUndefined	0.0142
reserved_room_typeD:mealUndefined	0.001693
reserved_room_typeE:mealUndefined	0.004125
reserved_room_typeF:mealUndefined	0.9766

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	22761 on 40500 degrees of freedom
Residual deviance:	14664 on 40423 degrees of freedom