Data Mining --- Exercise 2

Shuyan & Yangxi & Chen

Problem 1: Visualization

1. **Show average boardings grouped by hour of the day, day of week, and month:**

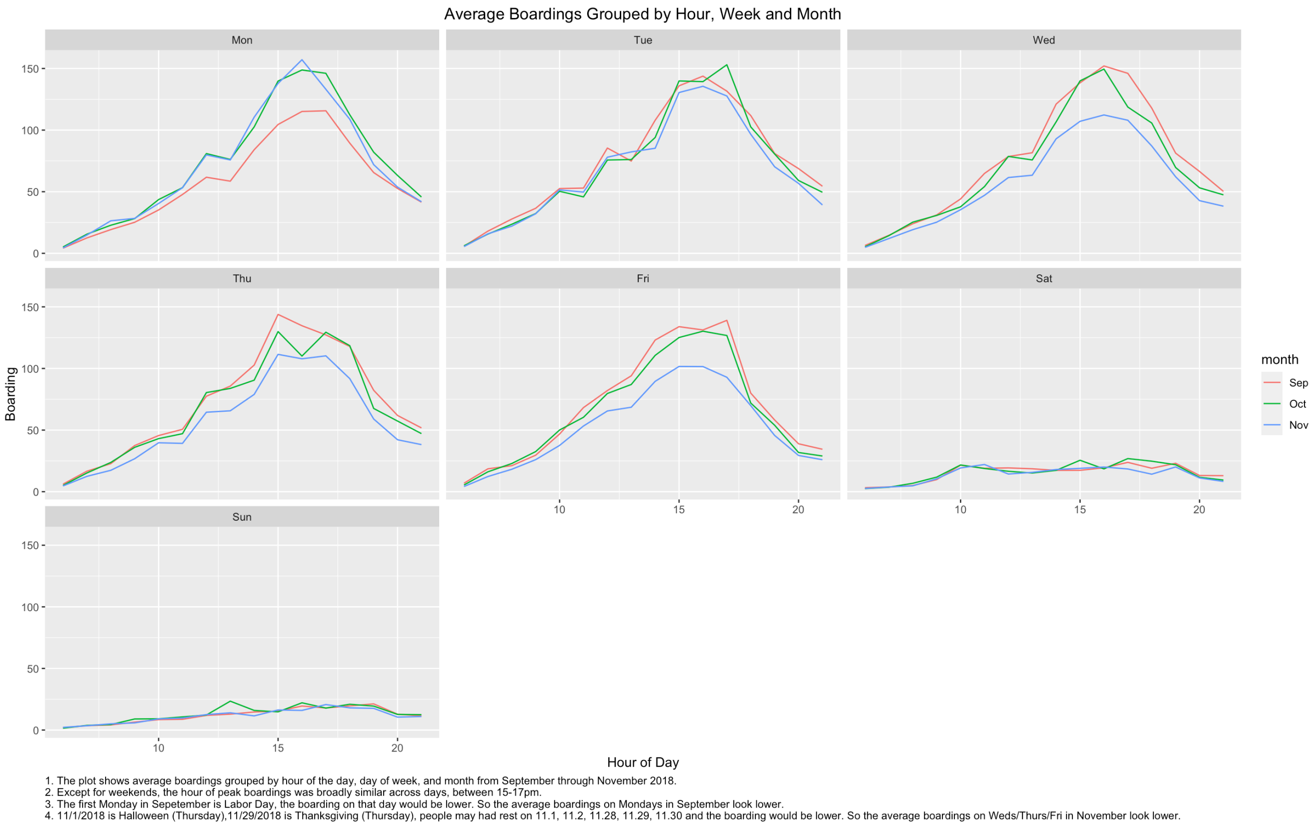


Figure 1: Average Boardings

2. **Show the relation between boardings and temperature:**

Figure 2: Relation between Boardings and Temperature

Problem 2: Saratoga House Prices

In this problem, we are going to choose the best model whose performance is better than the medium (baseline)model in predicting Saratoga house prices.

Medium Model: price ~ lotSize + age + livingArea + pctCollege + bedrooms + fireplaces+bathrooms + rooms + heating + fuel + centralAir

1. **The Best Linear Model**

Use forward selection method and stepwise selection to automatically select a model with the lowest AIC.

Baseline Model 1 (Medium Model)
price = lotSize + age + livingArea + pctCollege + bedrooms + fireplaces+bathrooms + rooms + heating + fuel + centralAir
Best Model 1 (Forward Selected Model)
price = livingArea + centralAir + bathrooms + bedrooms + fuel + lotSize + rooms + pctCollege + livingArea:centralAir + livingArea:fuel + centralAir:fuel + centralAir:bedrooms + centralAir:bathrooms + fuel:lotSize + bedrooms:fuel + centralAir:rooms + fuel:pctCollege + livingArea:rooms + livingArea:bedroomsr
Best Model 2 (Stepwise Selected Model
price = lotSize + age + livingArea + pctCollege + bedrooms + fireplaces + bathrooms + rooms + heating + fuel + centralAir + livingArea:centralAir + age:pctCollege + livingArea:fuel + age:fuel + bedrooms:centralAir + pctCollege:fireplaces + pctCollege:bathrooms + fuel:centralAir + livingArea:rooms + livingArea:bedrooms + pctCollege:fuel + lotSize:fireplaces + age:heating + lotSize:bedrooms + rooms:heating + rooms:fuel + bathrooms:centralAir + livingArea:fireplaces + bedrooms:fireplaces + fireplaces:centralAir + lotSize:livingArea

Baseline Model 2 (Medium Model with LandVaule)
price = landValue + lotSize + age + livingArea + pctCollege + bedrooms + fireplaces+bathrooms + rooms + heating + fuel + centralAir
Best Model 3 (Forward Selected Model)
price = livingArea + landValue + bathrooms + centralAir + lotSize + bedrooms + rooms + livingArea:centralAir + livingArea:landValue + livingArea:bathrooms + livingArea:lotSize + centralAir:bedrooms + centralAir:lotSize + landValue:lotSize + bathrooms:lotSize + centralAir:rooms
Best Model 4 (Stepwise Selected Model
price = landValue + lotSize + age + livingArea + pctCollege + bedrooms + fireplaces + bathrooms + rooms + heating + fuel + centralAir + landValue:age + livingArea:centralAir + livingArea:fuel + fuel:centralAir + landValue:lotSize + livingArea:fireplaces + landValue:fireplaces + age:centralAir + landValue:bathrooms + landValue:bedrooms + landValue:pctCollege + pctCollege:fireplaces + landValue:livingArea + age:bedrooms + lotSize:age + bedrooms:fireplaces + pctCollege:bedrooms + lotSize:fuel + lotSize:bathrooms + rooms:heating + livingArea:bedrooms

These six models are measured by the average out-of-sample RMSE. We average the performance of six models over 100 train/test splits by getting out-of-sample RMSE of each model. The following table shows average RMSE of six models:

Model	RMSE
Baseline Model 1	66503.29
Forward Model	64654.81
Stepwise Model	64251.13
Baseline Model 2	60284.24
Forward_Model_landValue	59643.77
Stepwise_Model_landValue	59623.12

According to the table, Stepwise_Model_landValue has the lower out-of -sample mean-squared error, which is 59623.12, and the average RMSE of the baseline model 1 is around 66503.29. So the Stepwise_Model_landValue makes an improvement. The regression result is:

	Stepwise_Model_landValue		
Predictors	Coefficients	CI	p
(Intercept)	72204.10	-10146.92 – 154555.12	0.086
landValue	0.06	-0.83 – 0.94	0.897
lotSize	20520.35	-6337.81 – 47378.51	0.134
age	-609.66	-1159.44 – -59.88	**0.030**

livingArea	81.99	50.69 – 113.29	**<0.001**
pctCollege	-1401.97	-2735.87 – -68.06	**0.039**
bedrooms	-25141.06	-50768.01 – 485.90	0.054
fireplaces	65049.08	23542.03 – 106556.13	**0.002**
bathrooms	9718.40	-1513.99 – 20950.80	0.090
rooms	3406.20	936.15 – 5876.24	**0.007**
heating [hot water/steam]	21479.99	-7445.01 – 50404.99	0.145
heating [electric]	22505.59	-19968.00 – 64979.18	0.299
fuel [electric]	-27785.02	-76578.25 – 21008.20	0.264
fuel [oil]	71468.01	29454.97 – 113481.05	**0.001**
centralAir [No]	15726.13	-12608.35 – 44060.62	0.276
landValue * age	0.00	0.00 – 0.01	**0.011**
livingArea * centralAir [No]	-17.03	-30.29 – -3.76	**0.012**
livingArea * fuel [electric]	7.52	-16.26 – 31.29	0.535
livingArea * fuel [oil]	-27.47	-44.47 – -10.46	**0.002**
fuel [electric] * centralAir [No]	14810.77	-6258.59 – 35880.13	0.168
fuel [oil] * centralAir [No]	-32103.83	-59712.35 – -4495.31	**0.023**
landValue * lotSize	-0.18	-0.39 – 0.04	0.104
livingArea * fireplaces	22.58	12.66 – 32.50	**<0.001**
landValue * fireplaces	-0.34	-0.54 – -0.14	**0.001**
age * centralAir [No]	411.12	5.65 – 816.59	**0.047**
landValue * bathrooms	0.42	0.21 – 0.63	**<0.001**
landValue * bedrooms	-0.08	-0.23 – 0.07	0.290
landValue * pctCollege	0.02	0.01 – 0.03	**0.003**
pctCollege * fireplaces	-926.31	-1541.38 – -311.25	**0.003**
landValue * livingArea	-0.00	-0.00 – -0.00	**0.001**
age * bedrooms	33.16	-83.45 – 149.77	0.577
lotSize * age	-155.03	-331.45 – 21.38	0.085
bedrooms * fireplaces	-12133.50	-21547.10 – -2719.91	**0.012**
pctCollege * bedrooms	451.29	12.47 – 890.10	**0.044**
lotSize * fuel [electric]	-4819.86	-23721.04 – 14081.31	0.617
lotSize * fuel [oil]	1153.79	-13254.30 – 15561.89	0.875
lotSize * bathrooms	-1777.83	-11442.73 – 7887.08	0.718

rooms * heating [hot water/steam]	-4213.81	-8068.36 – -359.27	**0.032**
rooms * heating [electric]	-3641.92	-8688.42 – 1404.57	0.157
livingArea * bedrooms	-0.09	-7.14 – 6.96	0.979

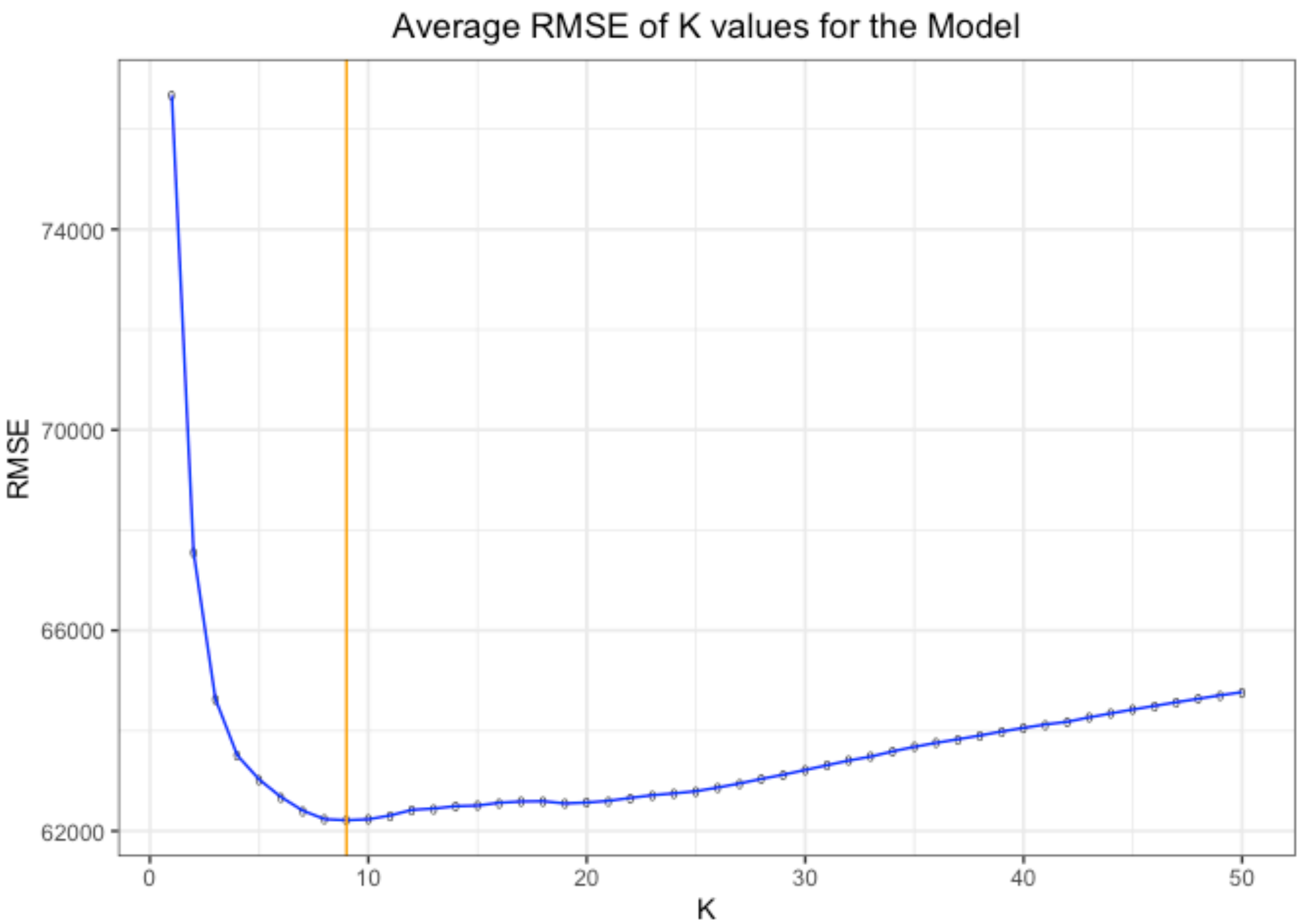
From the regression, we can conclude price-modeling strategies for a local taxing authority.

Firstly, the land value is the most significant factor of predicting house price because when adding this factor the RMSE dropped sharply. The higher the land value,the higher the house price.

Secondly, the lot size, bedrooms, fireplaces, heating , fuel, central air also affect the house price mostly. If there's no central air and more bedrooms, the house price will decrease. In addition, the larger the lot size, the higher the house price. Also, the availability of fireplaces also has a significant impact on home prices. Heating is important, no matter the type of heating. And comparing fuel with electric, fuel with oil will increase the house price.

Thirdly, the interaction between the type of fuel and central air is important. If without centrl air, fuel with electric will increase price comparing to the fuel with oil. And people tend to choose the bedrooms with fireplaces.

2. **Build the best K-nearest-neighbor regression model for price**

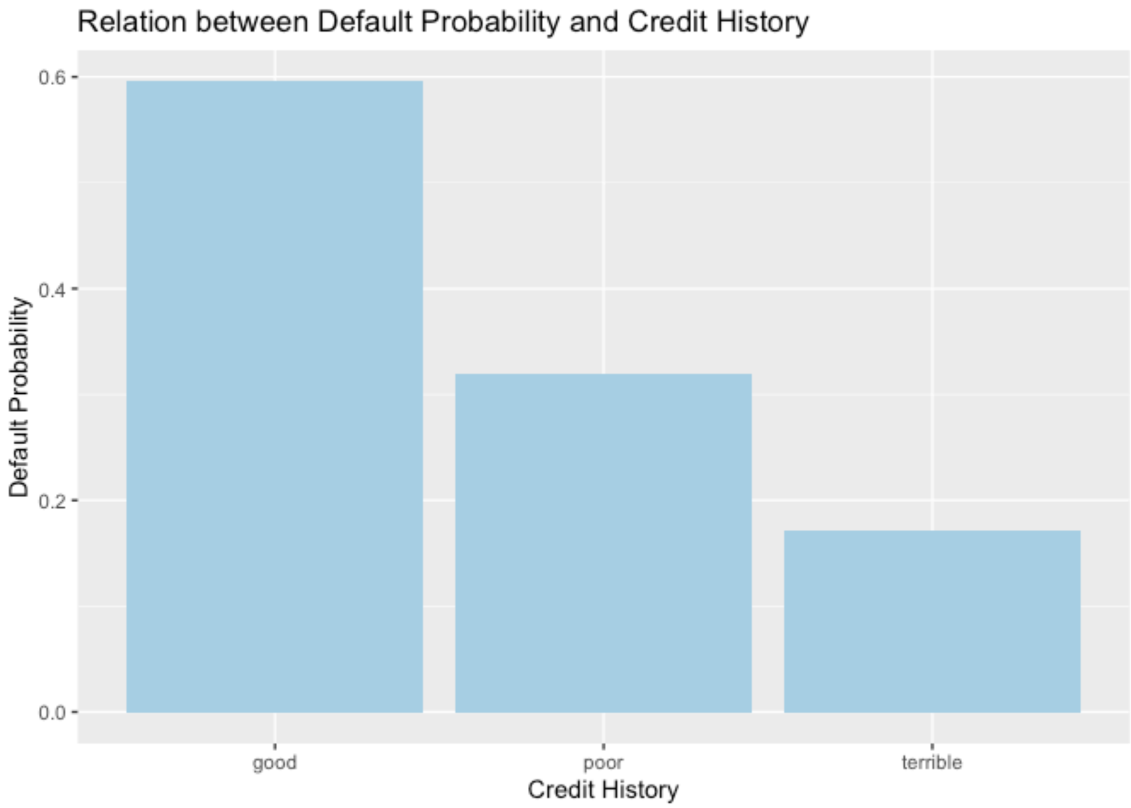


Across all 20 folds of the val set, from the plot above, we can see that the min average RMSE getting from KNN model is about 62000 with optimal K, which is about 7-9. The RMSE from KNN model is higher than the lowest RMSE from linear model that we talk above.

Thus, linear model does better at achieving lower out-of-sample mean-squared error.

Problem 3: Classification and Retrospective Sampling

1. **Bar plot of default probability by credit history**



2. **Build a logistic regression model for predicting default probability**

```
##
## Call:  glm(formula = Default ~ duration + amount + installment + age + history +
purpose + foreign, family = "binomial", data = german_credit)
##
## Coefficients:
##         (Intercept)             duration               amount
##          -7.075e-01            2.526e-02            9.596e-05
##         installment                  age          historypoor
##           2.216e-01           -2.018e-02           -1.108e+00
##     historyterrible          purposeedu  purposegoods/repair
##          -1.885e+00            7.248e-01            1.049e-01
##       purposenewcar       purposeusedcar        foreigngerman
##           8.545e-01           -7.959e-01           -1.265e+00
##
## Degrees of Freedom: 999 Total (i.e. Null);  988 Residual
## Null Deviance:       1222
## Residual Deviance: 1070  AIC: 1094
```

From the bar plot and logistic regression model, we could see that the default probability is high for those with good credit history and low for those with terrible credit history, which is contrary to our perceptions, so this data set may not be a good source to set a predicting model.

Because this data was collected in a retrospective way, the defaults rare, and the bank sampled a set of loans that had defaulted for inclusion in the study, which resulted in a substantial oversampling of defaults, relative to a random sample of loans in the bank’s overall portfolio. Some set of loans don’t default so they are considered as low default group, whose default rate was underrepresented.

For bank's sampling scheme, we suggest that we can use bootstrap when selecting sample.

Problem 4: Children and Hotel Reservations

1. **Model Building**

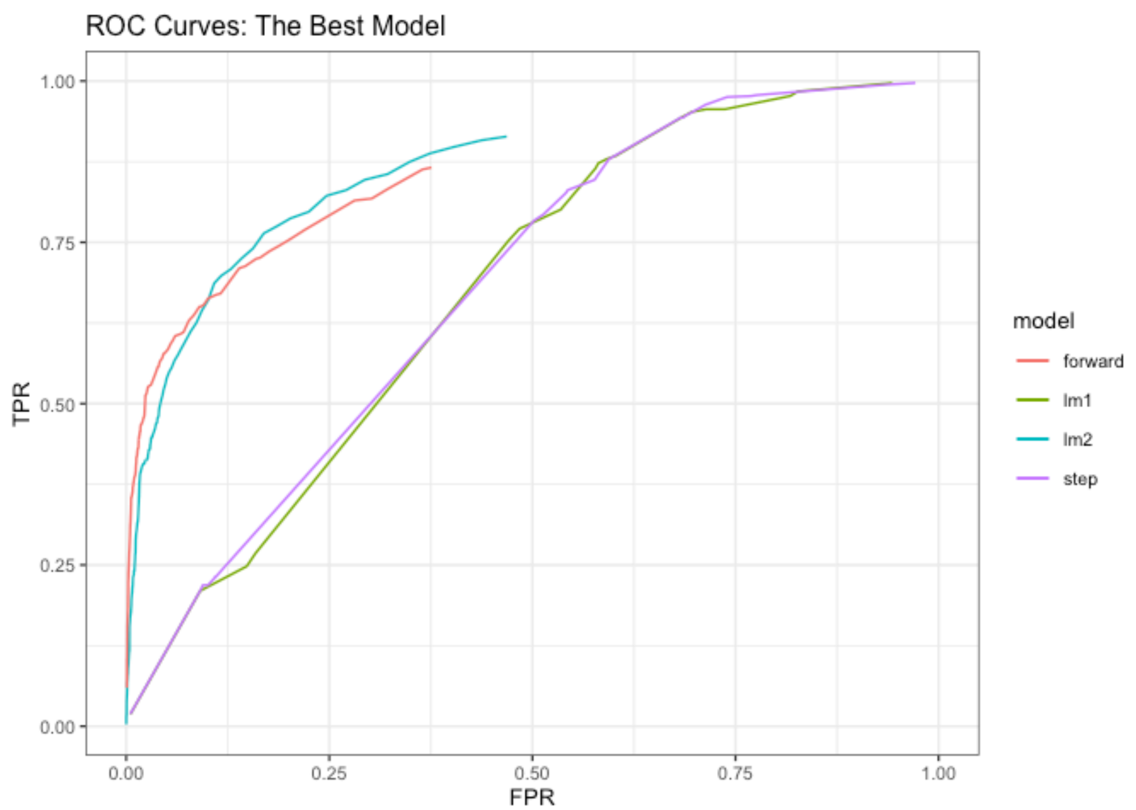
We will use TPR & FPR and deviance to check performance for the following models.

For baseline 1, this small model uses only the market_segment, adults, customer_type, and is_repeated_guest variables as features.

For baseline 2, this big model uses all the possible predictors except the arrival_date variable (main effects only) as features.

For our own model, we use forward selection and stepwise selection to find best linear model. For forward model, we choose adults, meal, market_segemnt,etc., a total of 11 features as main effect.

(1) Using TPR and FPR to check the performances for four models



From the ROC curve, we can see that the forward model is better off because the whole line relatively closes to the left corner.

(2) Compare four models in deviance

model	deviance
lm1	2622.312
lm2	1973.359
lm_forward	1743.190
lm_step	2609.227

From the table above, we can see that the forward model has the lowest deviance, which is 1743.190.

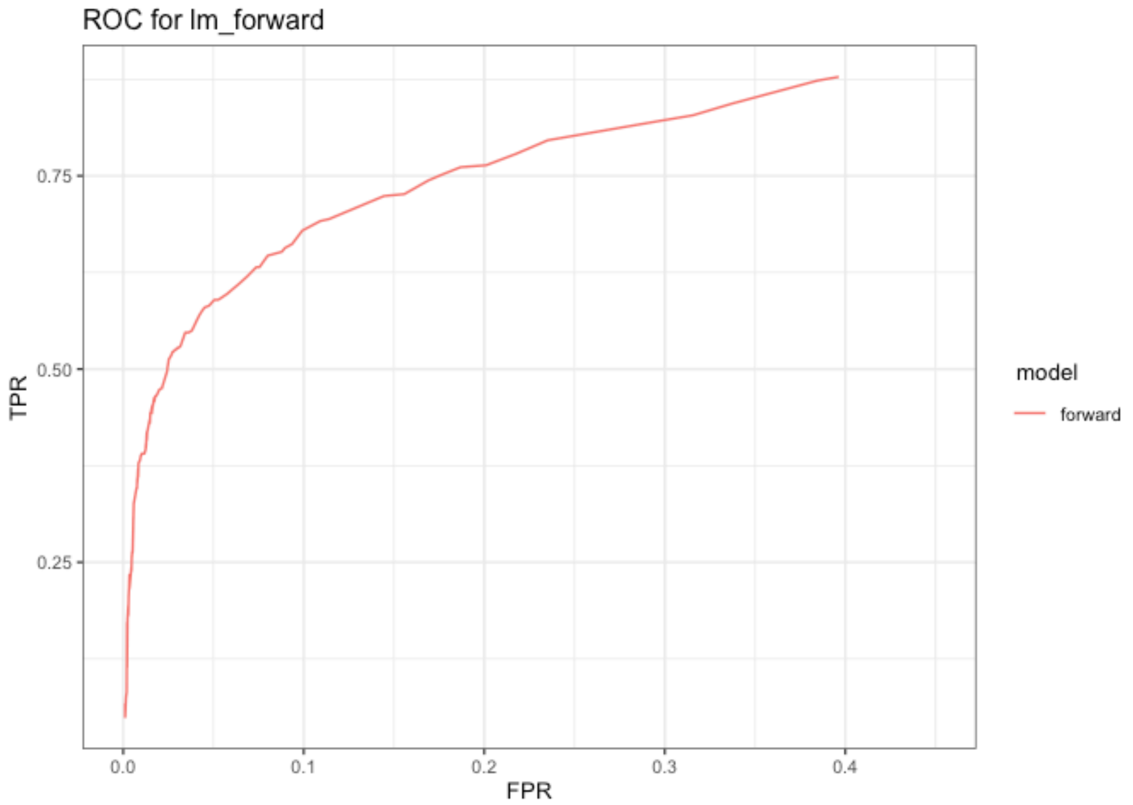
In conclusion. the forward model performs best out of sample.

Forward Model:

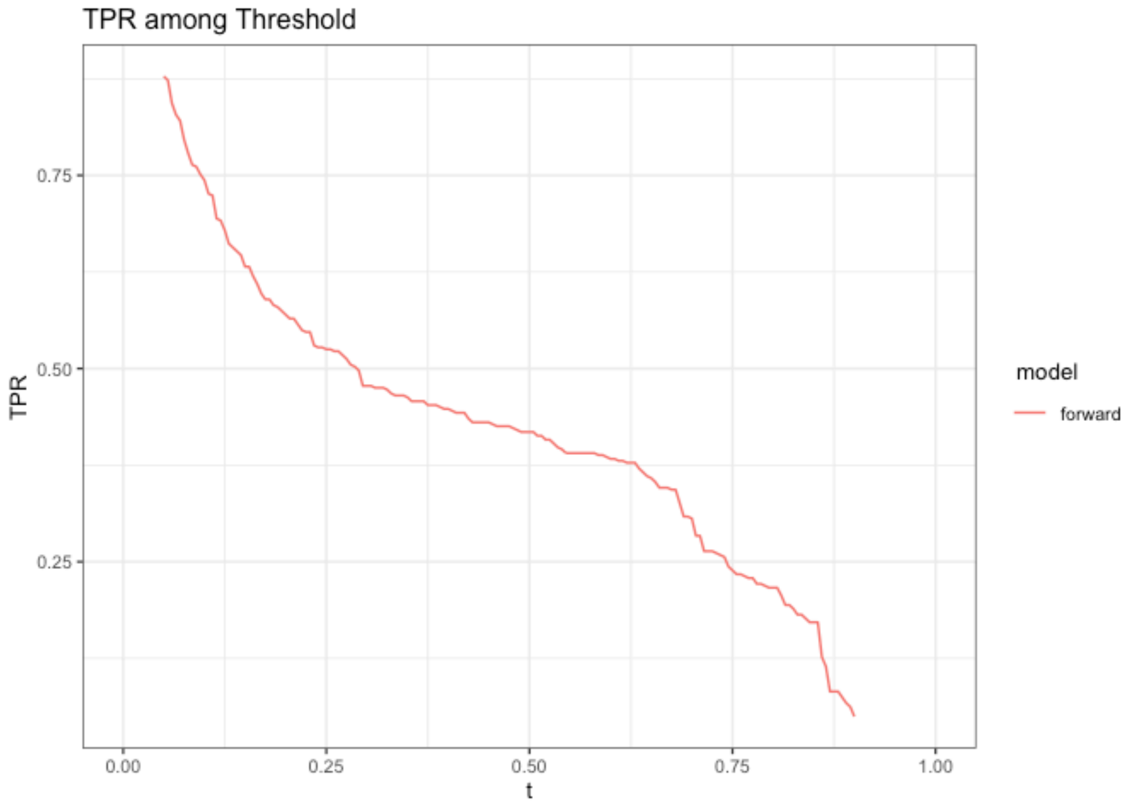
children = reserved_room_type + total_of_special_requests + assigned_room_type + hotel + market_segment + meal + adults + is_repeated_guest + required_car_parking_spaces + reserved_room_type:assigned_room_type + reserved_room_type:hotel + reserved_room_type:market_segment + assigned_room_type:hotel + total_of_special_requests:meal + hotel:market_segment +
reserved_room_type:meal + market_segment:meal + reserved_room_type:adults + assigned_room_type:market_segment + market_segment:adults + reserved_room_type:is_repeated_guest + total_of_special_requests:is_repeated_guest + total_of_special_requests:assigned_room_type + total_of_special_requests:market_segment + assigned_room_type:meal + total_of_special_requests:adults + meal:adults + reserved_room_type:total_of_special_requests + hotel:required_car_parking_spaces + meal:required_car_parking_spaces + reserved_room_type:required_car_parking_spaces + total_of_special_requests:required_car_parking_spaces + market_segment:required_car_parking_spaces + total_of_special_requests:hotel + meal:is_repeated_guest + hotel:adults

2. **Model Validation: Step 1**

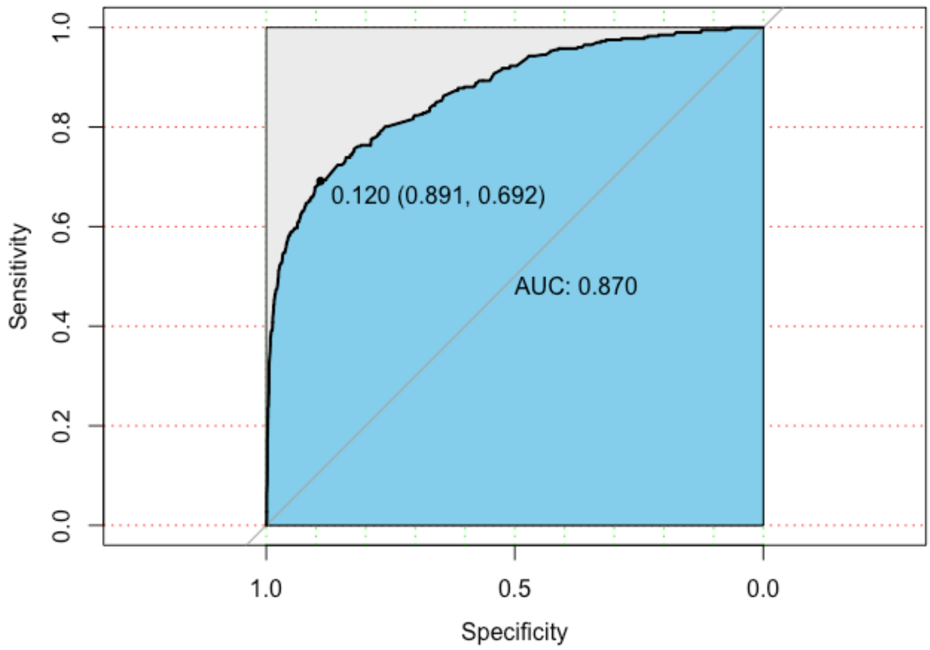
(1) ROC curve:



(2) Find the optimal threshold t:



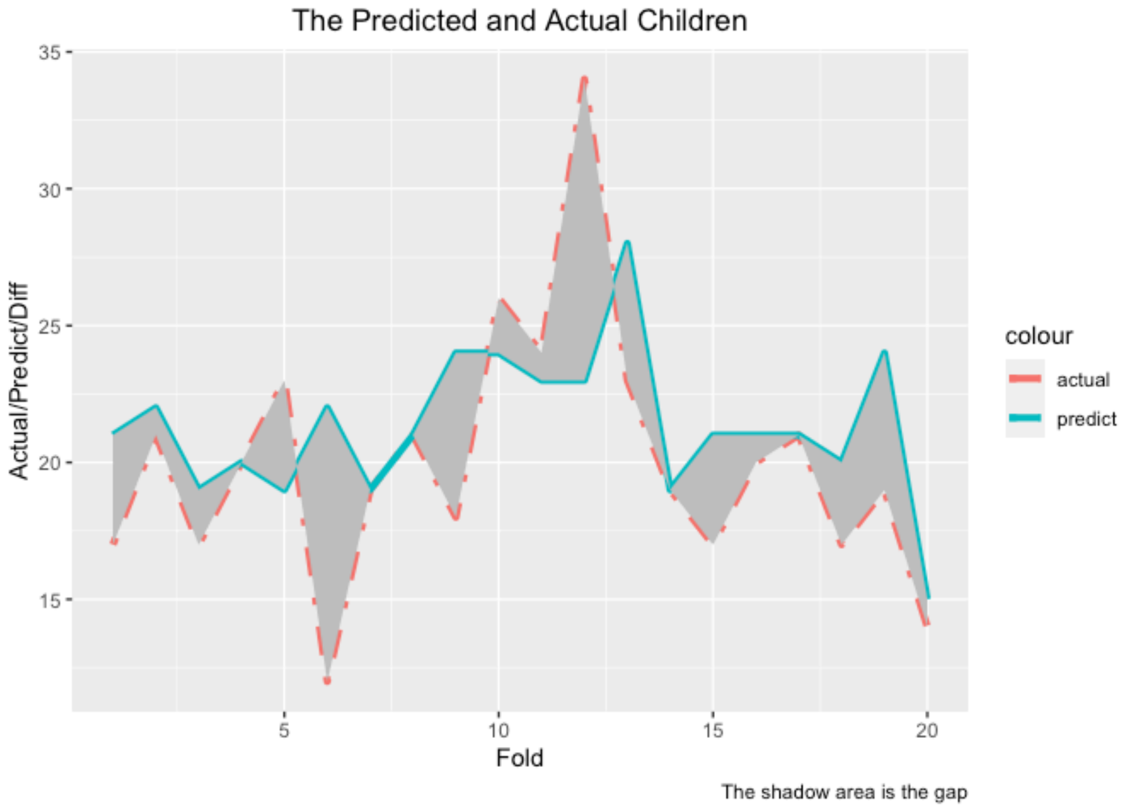
(3) Find feasible best t:



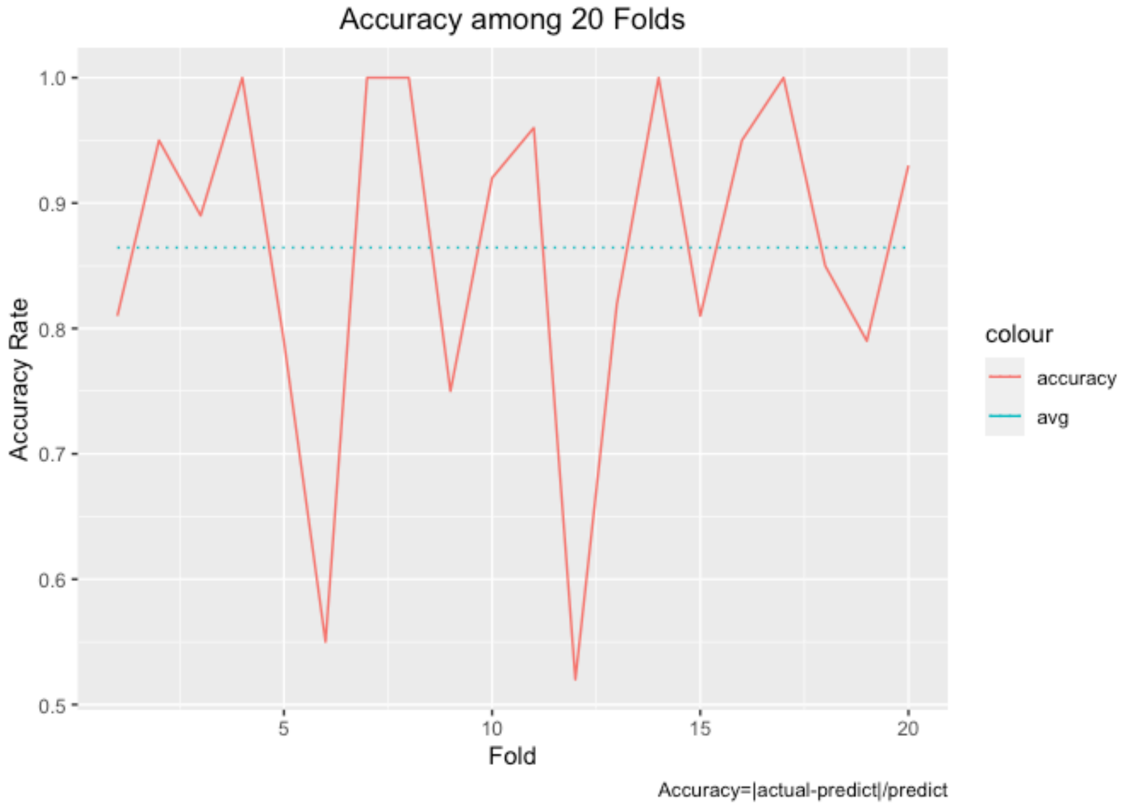
3. **Model Validation: Step 2**

The following show the actual and predicting of the total number of bookings with children in a group of 250 bookings. It also shows the difference between these two. By looking at the difference, we think the model does good because the predict numbers are close to the real.

Visualize the actual and predicting of the total number of bookings with children, and also the the difference between them:



Show the accuracy in different folds:



Show the detail of actual and predicting of the total number of bookings with children, and their difference in each fold:

fold	predict	actual	diff	accuracy
1	21	17	-4	0.81
2	22	21	-1	0.95
3	19	17	-2	0.89
4	20	20	0	1.00
5	19	23	4	0.79
6	22	12	-10	0.55
7	19	19	0	1.00
8	21	21	0	1.00
9	24	18	-6	0.75
10	24	26	2	0.92
11	23	24	1	0.96
12	23	34	11	0.52
13	28	23	-5	0.82
14	19	19	0	1.00
15	21	17	-4	0.81
16	21	20	-1	0.95
17	21	21	0	1.00
18	20	17	-3	0.85
19	24	19	-5	0.79
20	15	14	-1	0.93

Show the accuracy of prediction:

```
## [1] 0.8645
```

In conclusion, the performance of forward model for prediction is outstanding, which holds 86.45% rate of accuracy.