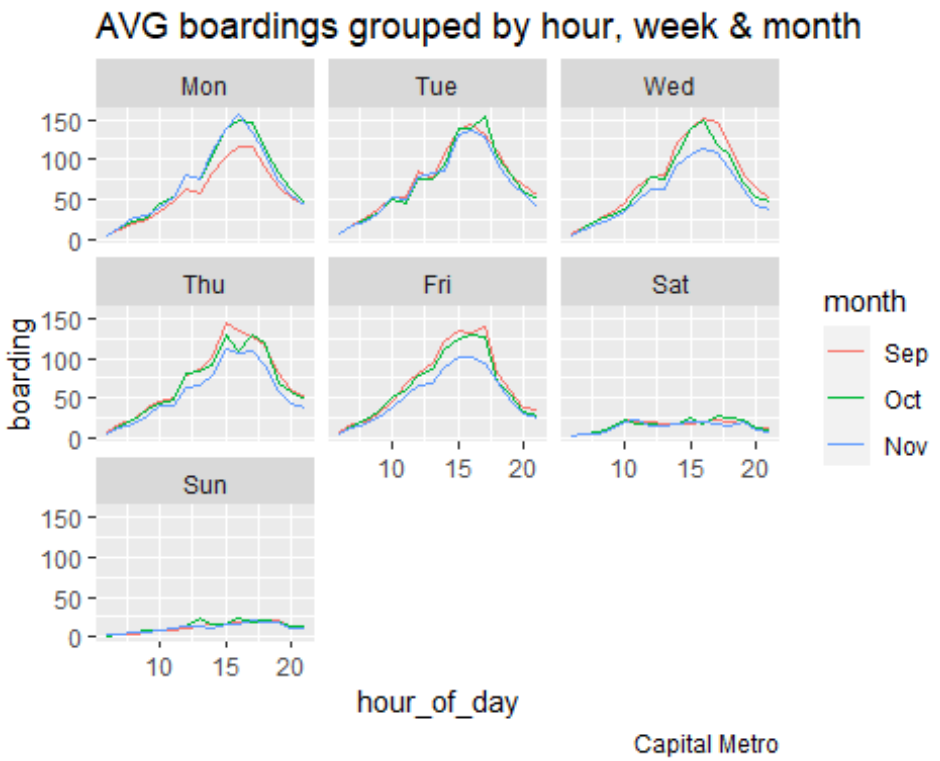
Excercise 2

Xiaohan Sun / Liyuan Zhang / Evelyn Cheng

2021/3/12

Problem 1: visualization

first figure + caption

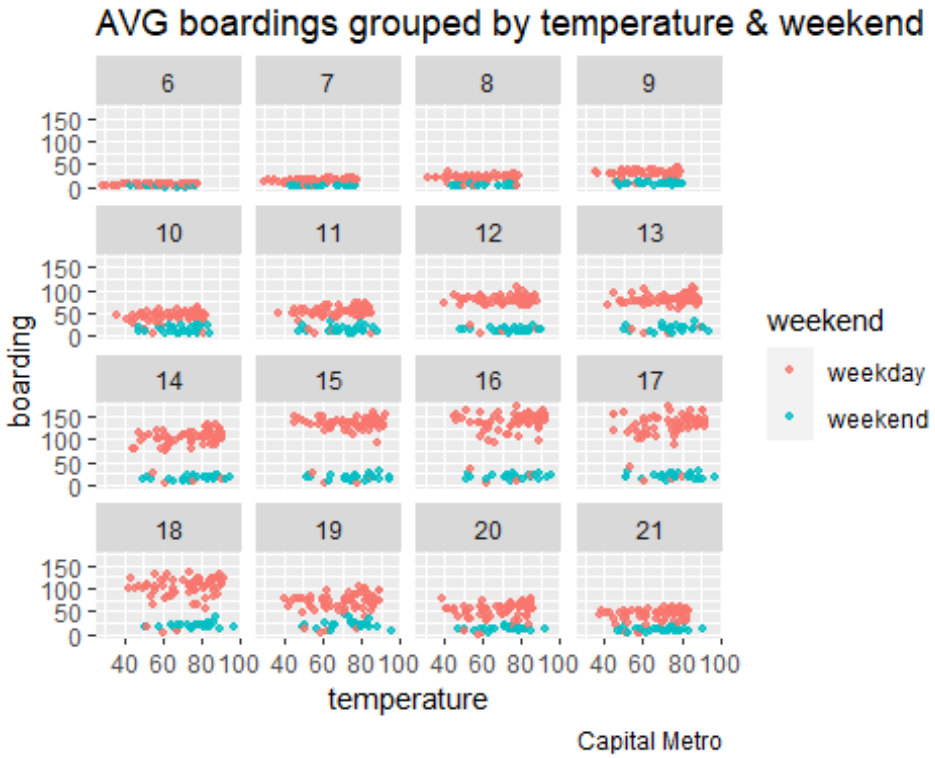


The hour of peak boardings change from day to day. During Monday to Friday which is the working day, the peak is 16-17 o'clock. During weekends, there is a smooth peak on 17-18 o'clock.

I reckon that the first Monday in September is Labor Day and there is one day off. So, average boardings on Mondays in September look lower.

Because there are two holidays in November, the 11th is a day off for Veterans Day. The second is Thanksgiving, the fourth Thursday in November, and there are two days off. Therefore, average boardings on Weds/Thurs/Fri in November look lower.

Second figure + caption



Temperature doesn't seem to have a noticeable effect on the number of UT students riding the bus. Since the points of the same color look like a straight line, there is no obvious slope.

problem 2: Saratoga house prices

Linear model

```
## [1] 66886.09

## [1] 85783.91

## [1] 65484.25

## [1] 66470.21

## lm(formula = price ~ livingArea + centralAir + bathrooms + fuel +
##     lotSize + bedrooms + rooms + livingArea:centralAir + livingArea:
bathrooms +
##     livingArea:fuel + bathrooms:bedrooms + bathrooms:fuel + fuel:lot
Size +
##     centralAir:bathrooms + livingArea:rooms + bedrooms:rooms +
##     centralAir:fuel, data = saratoga_train)
```

The best linear model I found is price = livingArea + centralAir + bathrooms + bedrooms +heating + lotSize + rooms + livingArea*centralAir + bathrooms*heating

+bedrooms*heating + livingArea*rooms + bedrooms*rooms + livingArea*lotSize +lotSize*rooms + centralAir*bedrooms. The RMSE is 65430.21, which is lower than the RMSE in professor's medium regression

KNN model

```
## lm(formula = price ~ livingArea + centralAir + bathrooms + bedrooms
+
##     lotSize + fuel + rooms, data = saratoga_train)
```

I will use the regression price=livingArea + centralAir + bathrooms + bedrooms +lotSize + rooms + heating to find the best RMSE in KNN model

```
## [1] 10
```

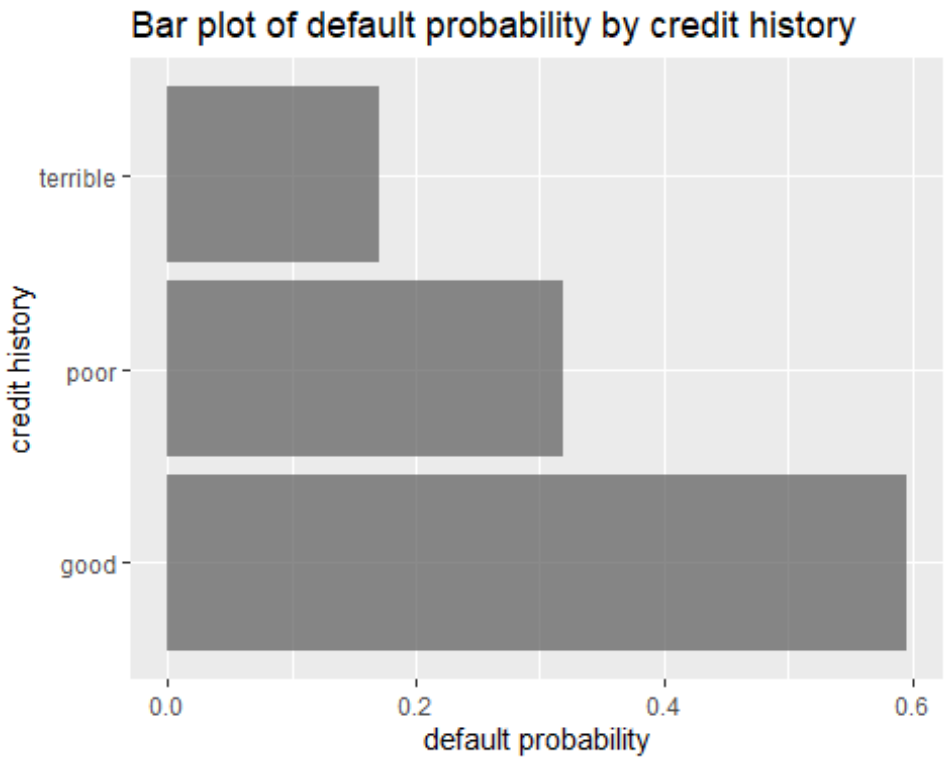
I will use this best K to find the RMSE

```
## [1] 64235.65
```

I think KNN model is much better than the linear model, since the RMSE in KNN model is greater than the linear model.

Problem 3: Classification and retrospective sampling

Make a bar plot of default probability by credit history



Build a logistic regression model for predicting default probability

```
##
## Call:
## glm(formula = Default ~ duration + amount + installment + age +
##     history + purpose + foreign, family = binomial, data = german_cr
edit)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3464  -0.8050  -0.5751   1.0250   2.4767
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -7.075e-01  4.726e-01  -1.497  0.13435
## duration             2.526e-02  8.100e-03   3.118  0.00182 **
## amount               9.596e-05  3.650e-05   2.629  0.00856 **
## installment          2.216e-01  7.626e-02   2.906  0.00366 **
## age                 -2.018e-02  7.224e-03  -2.794  0.00521 **
## historypoor         -1.108e+00  2.473e-01  -4.479 7.51e-06 ***
## historyterrible     -1.885e+00  2.822e-01  -6.679 2.41e-11 ***
## purposeedu           7.248e-01  3.707e-01   1.955  0.05058 .
## purposegoods/repair  1.049e-01  2.573e-01   0.408  0.68346
## purposenewcar        8.545e-01  2.773e-01   3.081  0.00206 **
## purposeusedcar      -7.959e-01  3.598e-01  -2.212  0.02694 *
## foreigngerman       -1.265e+00  5.773e-01  -2.191  0.02849 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1221.7  on 999  degrees of freedom
## Residual deviance: 1070.0  on 988  degrees of freedom
## AIC: 1094
##
## Number of Fisher Scoring iterations: 4
```

What do you notice about the history variable vis-a-vis predicting defaults? What do you think is going on here?

Banks provide high amount loans to people with good credit history , but without collateral. This is why people with terrible credit history default on their loans when they do this. Since defaults are rare, the bank conducted a sample survey of a group of defaulted loans. Banks try to match each default behavior with similar loan groups that have not defaulted, resulting in a large number of default over-sampling. In the graph we've made, the lower the historical credit of the borrower, the lower the probability of default.

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?

According to what I've done in this question, I reckon that this data set isn't appropriate for building a predictive model of defaults. Because there is a substantial oversampling of defaults. In my opinion, I recommend bank should reduce the sample of defaults. Using proportional sampling instead, it maybe more appropriate for predictive model of defaults.

Problem 4: Children and hotel reservations

Model building

Baseline 1

For baseline 1, the model is just contain 4 features with no interaction, and we use MSE to evaluate the out-of-sample performance.

MSE for baseline 1: 0.2678641.

Baseline 2

For baseline 2, this is a big model with all features and interactions except `arrival_date`.

MSE for baseline 2: 0.2244956.

Baseline 3

For baseline 3, there are two models that use different selection methods. One uses stepwise selection, the other uses forward selection.

For forward selection, we choose the features that we think is essential to the outcome such as `stays_in_weekend_nights`, `adults`, `meal`, `reserved_room_type`, etc.

MSE for model with stepwise selection: 0.2628751.

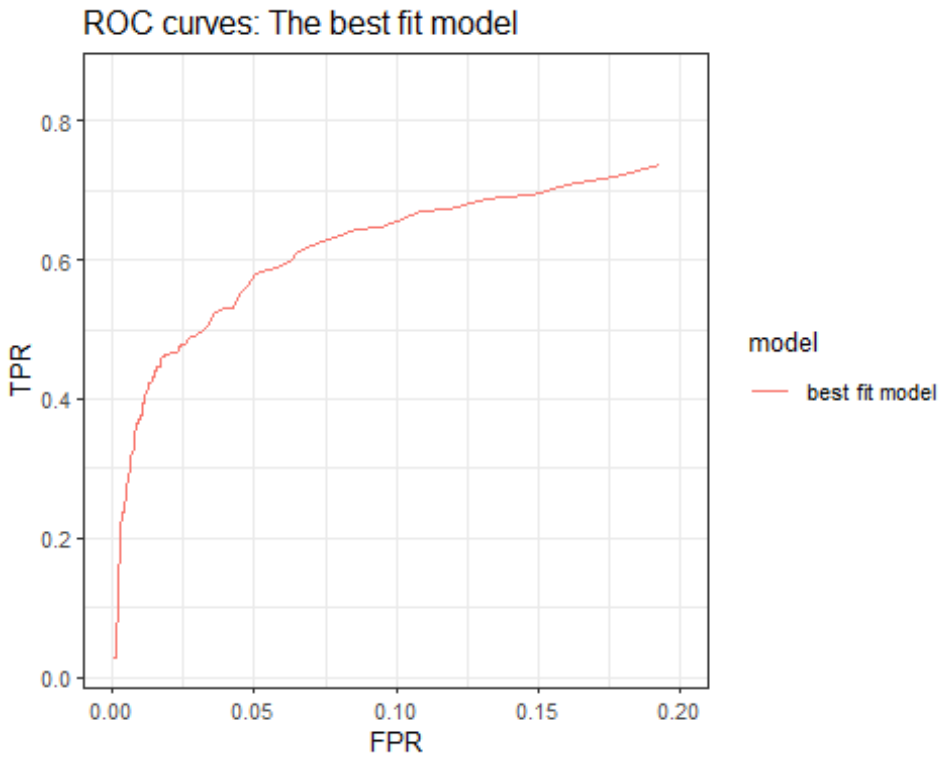
MSE for model with forward selection: 0.2173147.

As a result, the model with forward selection is better.

After comparing the MSE values of these models, we find that the model in baseline 3 with forward selection is the best.

Model validation: step 1

This is the ROC curve we produced for our best model. The data we use is `hotels_val`. And we just zoomed in for FPR between 0 and 0.2.



Model validation: step 2

The following table shows the actual and predict number of children, also the difference between these two. And we find that most of predict numbers is close to the real number. If we set that the difference that beyond 5 is false value, then the precision of this model is 80%.

```
## # A tibble: 20 x 4
##    fold_id Actual Predict  diff
##  *   <int>  <int>   <int> <int>
##  1       1     19      20    -1
##  2       2     22      23    -1
##  3       3     24      16     8
##  4       4     26      16    10
##  5       5     18      21    -3
##  6       6     19      11     8
##  7       7     18      13     5
##  8       8     22      17     5
##  9       9     31      18    13
## 10      10     14      18    -4
## 11      11     20      14     6
## 12      12     12      14    -2
## 13      13     20      19     1
## 14      14     17      12     5
## 15      15     18      20    -2
## 16      16     22      15     7
## 17      17     18      19    -1
```

```
## 18      18      18      16      2
## 19      19      21      20      1
## 20      20      23      16      7
```