Exercise2

Zonghao Li

2021/3/11

Problem1-Visualization

At first, we should recode the categorical variables in sensible, rather than alphabetical, order.

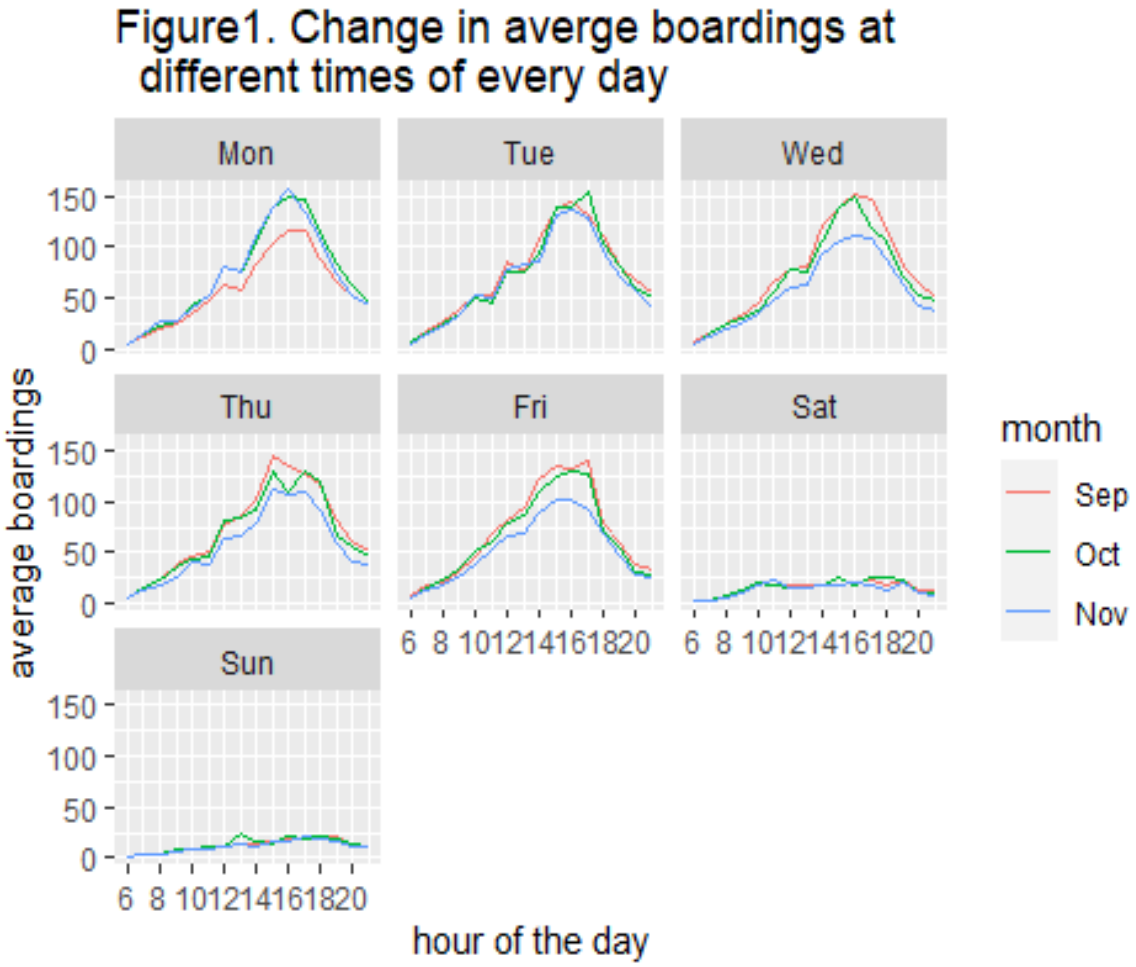


Figure1 describes the change in averge boardings at different times of every day in the week, showing the different trends of conditions in September, Ocotober and November. When checking the dataset, I found that there was no data before 6am or after 9pm, so the data from 6am to 9pm was selected.

(1) From Figure1, we can learn that the hour of peak boardings is broadly similar across days in the week. In particular, the boardings significantly decrease during the weekends due to not working. Furthermore, the boardings change smoothly during one day and there is no obvious peak boardings.

(2) Then we can see that average boardings on Mondays in September look lower, compared to other days and months. I think this is caused by the Labor Holiday, which is on the first Monday in September. During that day, people will not take classes or works so the boardings falls down dramatically.

(3) Similarly, this figure also shows that average boardings on Weds/Thurs/Fri in November look lower. I think during that period, people will prepare for some presentations or exams since that period is close to the end of the semester.

Figure2 shows the relationship between boardings and temperature at different times of the day, and it is grouped by weekend or not. In general, holding weekend and hours of the day fixed, this figure conveys that temperature does not have a significant effect on the number of UT students riding the bus. In particular, in some hours, there is a slight upward trend which means as the temperature gores up, the number of boardings will increase a bit, ignoring several outliers.

Problem2-Saratoga house prices

In this problem, we will run a "horse race" (i.e. a model comparison exercise) between two model classes: linear models and KNN.

[linear model]

```
## [1] 65002.06

## [1] 66001.88

## [1] 62005.72
```

After comparing the values of RMSE of three models, finally we will choose model 3 as the best linear model. In this model, price is dependent variable, and independent variables contain lotSize, age, livingArea, bedrooms, fireplaces, bathrooms, rooms, heating, fuel, centralAir, livingArea×centralAir, livingArea×fuel, bathrooms×heating, age×fuel, livingArea×fireplaces, bedrooms×fireplaces, fireplaces×centralAir, fuel×centralAir, age×centralAir, rooms×heating, lotSize×fireplaces.

[knn model]

At first, we should normalize the variables before applying KNN. In order, the k-values are: k=1, k=2, k=3, k=6, k=10, k=20, k=30.

```
## [1] 79334.11

## [1] 71980.65

## [1] 70733.86

## [1] 67921.1

## [1] 67830.35

## [1] 67659.22

## [1] 68521.29
```

After trying several values of k, we can approximately get that the out-of-sample mean-squared error value of knn model is larger than the linear model in the first part. So for this project, I recommend using the linear model to predict market values better and set prices more properly.

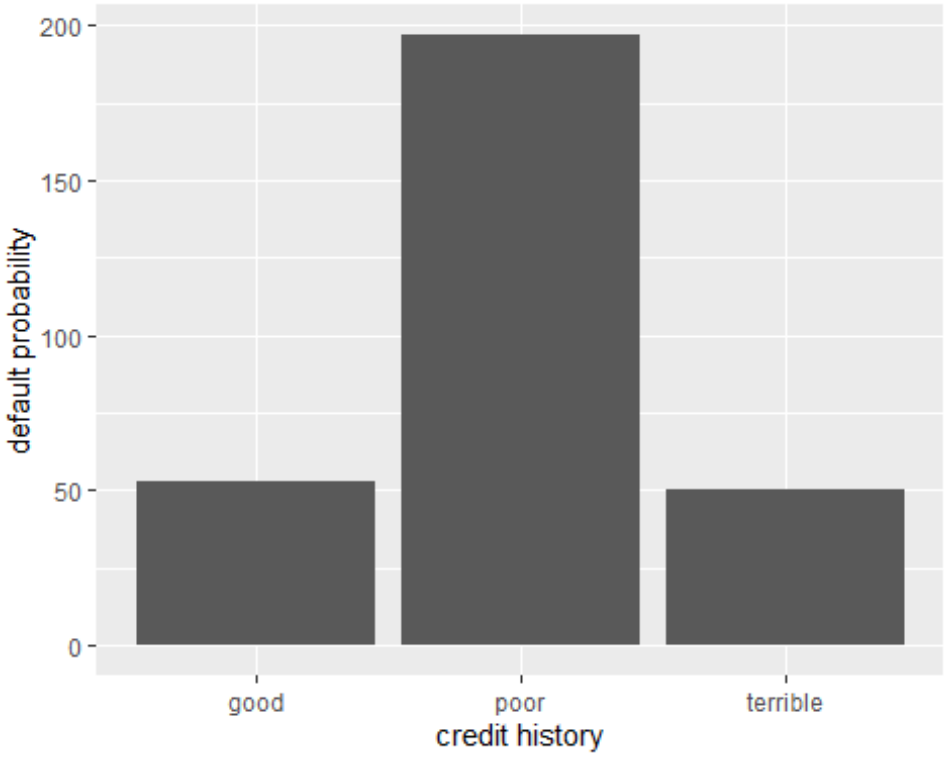
Problem3-Classification and retrospective sampling

In problem3, we will focus on helping the bank predict whether a borrower is likely to default on a loan.

```
##
## 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
```

We can notice that history variable has been divided into two variables–historypoor and historyterrible. And the estimated values are negative and gtatistically significant at 1% level, which show that the poor credit history would reduce the probability of default in fact. From my point of view, this result may caused by the different level of loans among groups of people with different credit history. The bank may send out some high-risk loan to people who have good credit history, but it may be more possible to make these people to default than people with poorer credit.



From the plot we can learn that people with poor credit history accounts for a large proportion, which means these people will more possible to default the loan. The probabilities of default among people with good credit or terrible credit are simliar and low.

```
##
## glm.pred   0   1
##     high 390 264
##     low  310  36
```

From the table above, we can calculate the probability of predicting correctly to be (390+36)/1000=0.426. This is a not too bad prediction probability, but the bank should take some ways to improve the predictive performance. For me, we can further use the knn method but its classification principle is different from the method above. Furthermore, random forest may be a possibly good way to make the prediction for bank.

Question4-Children and hotel reservations

The goal of this problem is to build a predictive model for whether a hotel booking will have children on it, since parents often enter the reservation exclusively for themselves and forget to include their children on the form when booking the hotel.

Model building

In this part, we need to compare several models below.

At first, split our 'hotels_dev' data set into training and testing sets.

[baseline1]

The value of RMSE of baseline1 is:

```
## [1] 0.2632775
```

[baseline2]

The value of RMSE of baseline2 is:

```
## [1] 0.27079
```

Next we will try to find a best linear model. The value of RMSE of this model is:

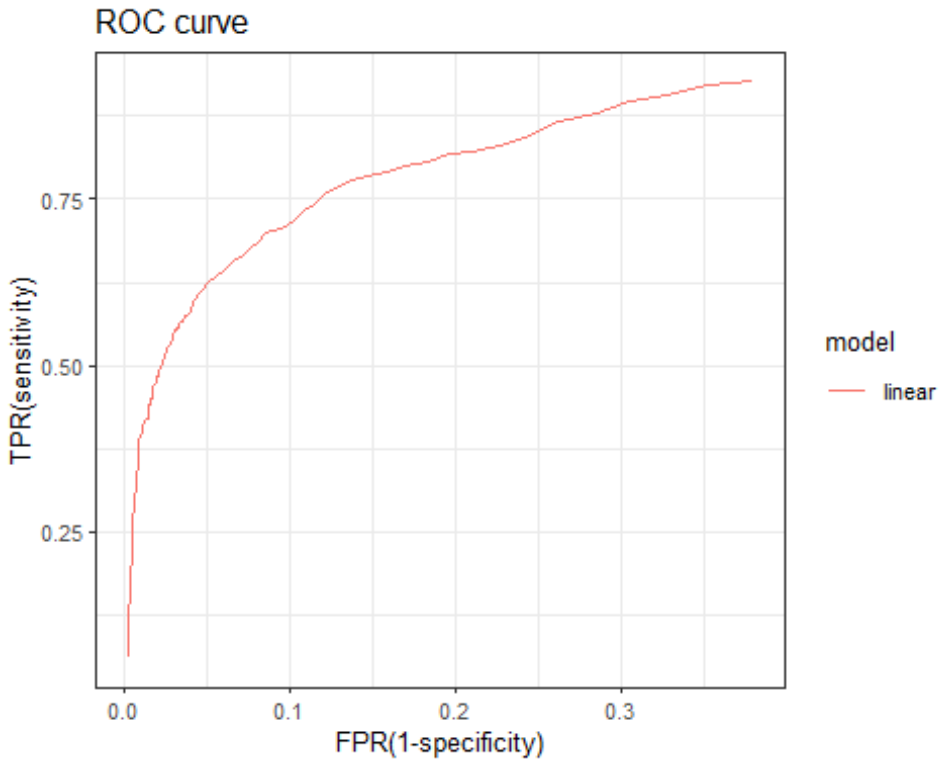
```
## Warning in predict.lm(model, data): prediction from a rank-deficient
 fit may be
## misleading

## [1] 0.2629362
```

From the model buildings and calculations above, we can learn that the best linear model in fact should be baseline2. So in the next section we will apply the baseline2 to analyze problems.

Model validation: step 1

Note: In two steps of model validation, we use only 'hotel_val' dataset.



The ROC curve is plotted above.

Model validation: step 2

This time we need to create 20 folds of 'hotels_val'.

```
## # A tibble: 20 x 3
##     fold  yhat     y
##   * <int> <int> <int>
##  1     1    18    19
##  2     2    20    19
##  3     3    28    22
##  4     4    34    22
##  5     5    33    22
##  6     6    18    12
##  7     7    30    24
##  8     8    29    22
##  9     9    21    14
## 10    10    22    23
## 11    11    26    21
## 12    12    28    19
## 13    13    24    21
## 14    14    22    16
## 15    15    27    29
```

```
## 16    16    20    17
## 17    17    25    14
## 18    18    29    21
## 19    19    23    25
## 20    20    27    20
```

From the table above, in particular, the differences between yhat and y range from 1 to 11. Some of the predicted values of number of children may differ a little greatly from the actural values, but overall, this prediction is fine.