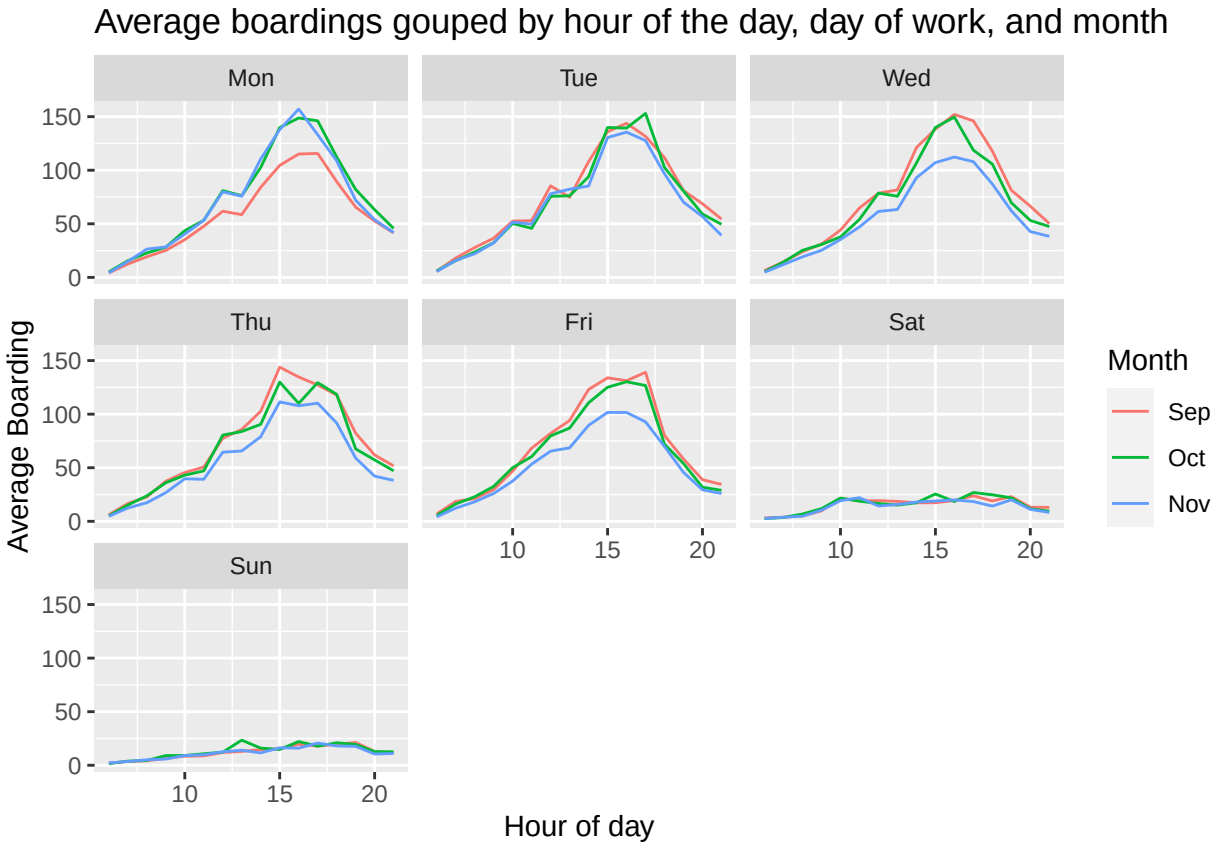
HW2 New revision

Zhiqian Chen, Qihang Liang, Yi Zeng

3/26/2021

##Problem 1: visualization

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

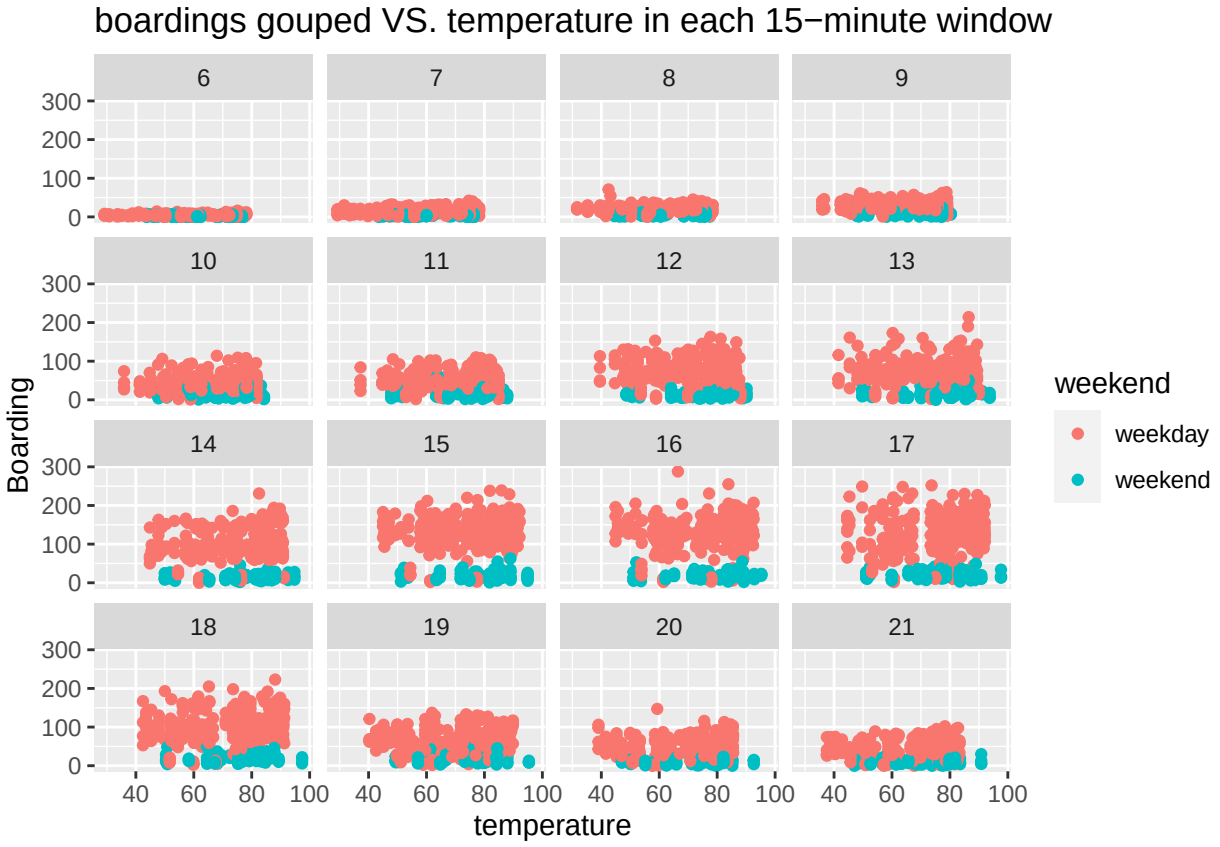


The graph whows that the it broadly similar across days for weekday or weekend. But there is a big bifferent between weekday and weekend.

The average boardings on Mondays in September look lower. The reason may be the hot weather in September, and the students don't want to board compare to other month. The other reason is that there is a holiday on monday in september. The average boardings on Weds/Thurs/Fri in November look lower. This may because the cold weather in November, students don't want to board compare to other month. The other reason is that the thanksgiving day is in November. Therefore the bordings is lower on Weds/Thurs/Fri in November.

#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.



When we hold hour of day and weekend status constant, temperature seems to have a no effect on the number of UT students riding the bus. But there is a big effect for bordings depends on weekday or weekend.

Problem 2

```
## [1] 69212.58
```

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

```
## [1] 152353.1
```

```
## [1] 67108.4
```

```
## [1] 73561.2
```

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

The best linear model I found is feature of 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 60363.84, which is lower than the RMSE in professor's medium regression. I will use the regression price=livingArea + centralAir + bathrooms + bedrooms +lotSize + rooms + heating to find the best RMSE in KNN model.

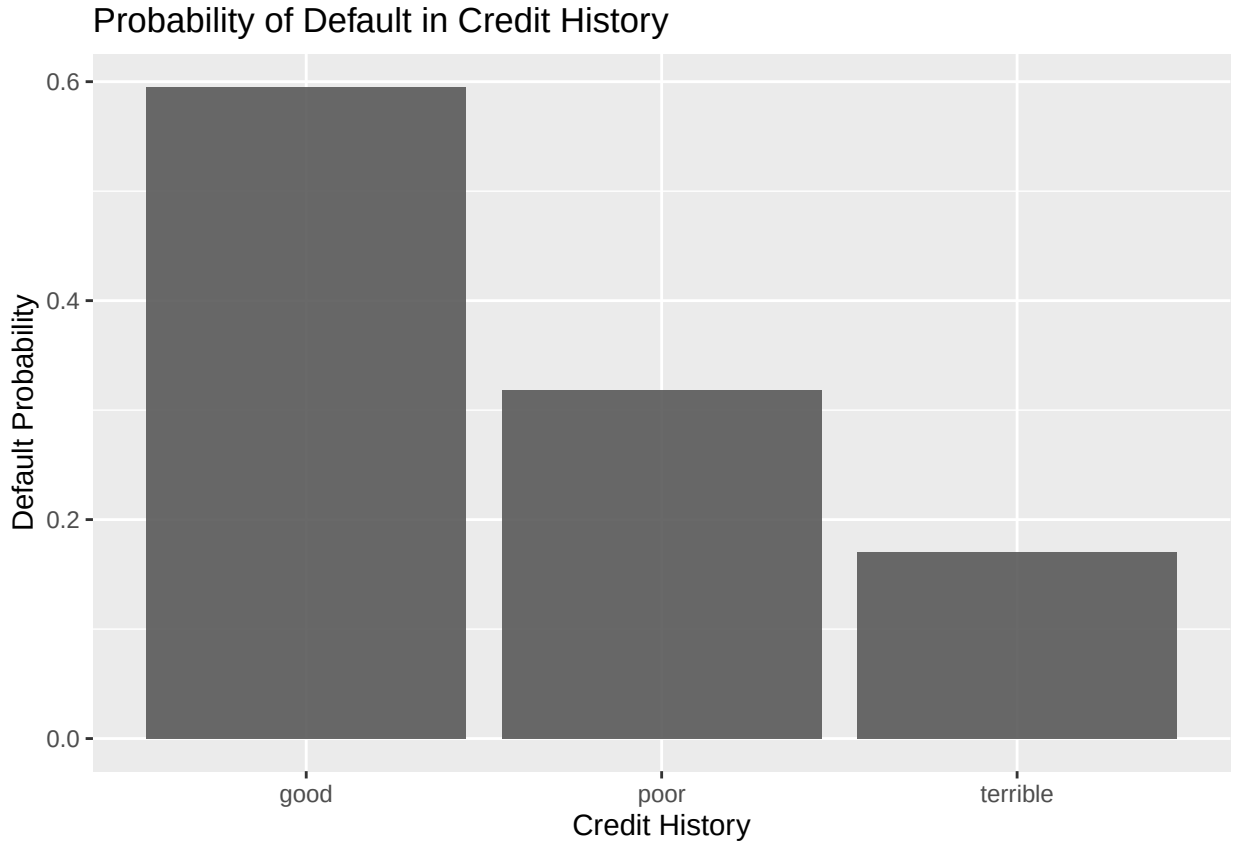
```
## [1] 10
```

So, here we get the k_best is 10, then we will use this k to generate the RMSE of the model that we find in the last step which is price=livingArea + centralAir + bathrooms + bedrooms +lotSize + rooms + heating.

```
## [1] 68908.45
```

Here, we find that the RMSE of the KNN model is 61127 which is larger than the linear model, so we can conclude that the KNN model is better fit the data than linear model.

Problem 3




```
##
## Call:
## glm(formula = Default ~ duration + amount + installment + age +
##     history + purpose + foreign, family = binomial, data = german_credit)
##
## 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
```

1)Banks provide a large number of loans to people with good credit records, but rarely provide loans to people with poor credit records. Since defaults rarely occur, the bank conducted a sample survey of a group of default loans. Banks try to match each default behavior with similar loan groups that have not yet defaulted, resulting in a large number of default over-sampling. According to the chart produced, the lower the historical credit of the borrower, the lower the probability of default. 2)I think this data set is not suitable for constructing the default prediction model because there are a large number of default values oversampling. In my opinion, I suggest that banks should reduce the default sample and increase the use of proportional sampling, which may be more suitable for default prediction models.

Problem 4: Children and hotel reservations

#Model Building

#1.baseline model 1: A small model that uses only the market segment, adults, customer type and repeated guest variables as features

```
## [1] 0.265868
```

For baseline model 1, the Root mean squared error is 0.27.

#2.baseline 2: a big model that uses all the possible predictors expect the arrival data variable

```
## [1] 0.2341166
```

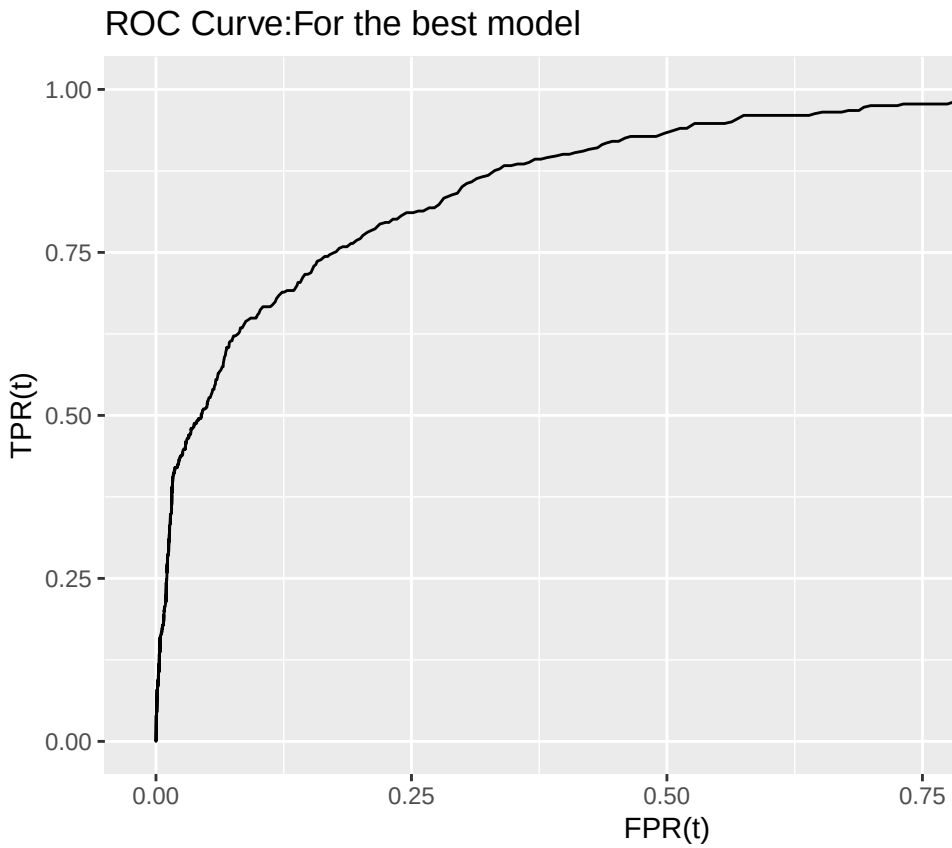
For baseline model 2, the Root mean squared error is 0.23.

#3.best linear model, as the question stated, hotel booking with children on it may affect hotel's restaurant (variable meal as feature), and including many engineered features, like previous booking canceled and customer type. Also, there may be some nonlinear relationship with age and the stays in weekend nights, so I add the the squared term in the model.

```
## [1] 0.2345737
```

For the best model I build, the Root mean squared error is 0.23 which is very close to baseline model 2. For the out-of-sample performance,the linear model I built has the minimum RMSE, and baseline model 1 has the larger RMSE. Also, the model I built has largest adjusted R-squared, so I will select bestmodel as the best model.

#Model Validation: Step 1 In model validation step 1, we plot the ROC curves for our best model with data in



hotels_val, with threshold vary from [0, 1]

#Model Validation: Step 2

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.16849  0.01017  0.05232  0.08450  0.10164  0.93795
```

```
## # A tibble: 20 x 4
##    fold_id True_Probability Predicted_Probability Residual
##  *   <int>            <int>                <int>    <int>
##  1       1               16                   15        1
##  2       2               22                   15        7
##  3       3               26                   10       16
```

```
##  4        4                14                9         5
##  5        5                25               20         5
##  6        6                24               14        10
##  7        7                16                9         7
##  8        8                19               13         6
##  9        9                18                9         9
## 10       10                19               19         0
## 11       11                18               15         3
## 12       12                25               18         7
## 13       13                20               12         8
## 14       14                18               11         7
## 15       15                20               12         8
## 16       16                19               17         2
## 17       17                16               16         0
## 18       18                19               15         4
## 19       19                23               12        11
## 20       20                25               16         9
```

In model validation step 2, we create 20 folds of hotels_Val and each fold have about 250 bookings in it,for each fold predict whether each booking will have children on it, sum up the predicted probability for all the bookings in the fold and compare this "expected" number of bookings with children versus the actual number of bookings with children in that fold.The error of my model do at predicting is shown above, there is a 0, and there are some large residual. However, I think the model do a great job.