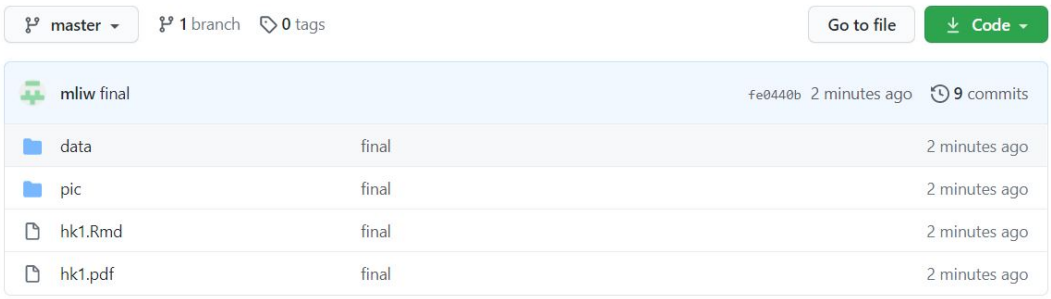
ECO395M STAT LEARNING Homework 2*

Mingwei Li, Xinyu Leng, Hongjin Long

Abstract

This document is the second homework of ECO395M STAT LEARNING.



*Mingwei Li, Xinyu Leng and Hongjin Long are master students of economics, The University of Texas at Austin

Contents

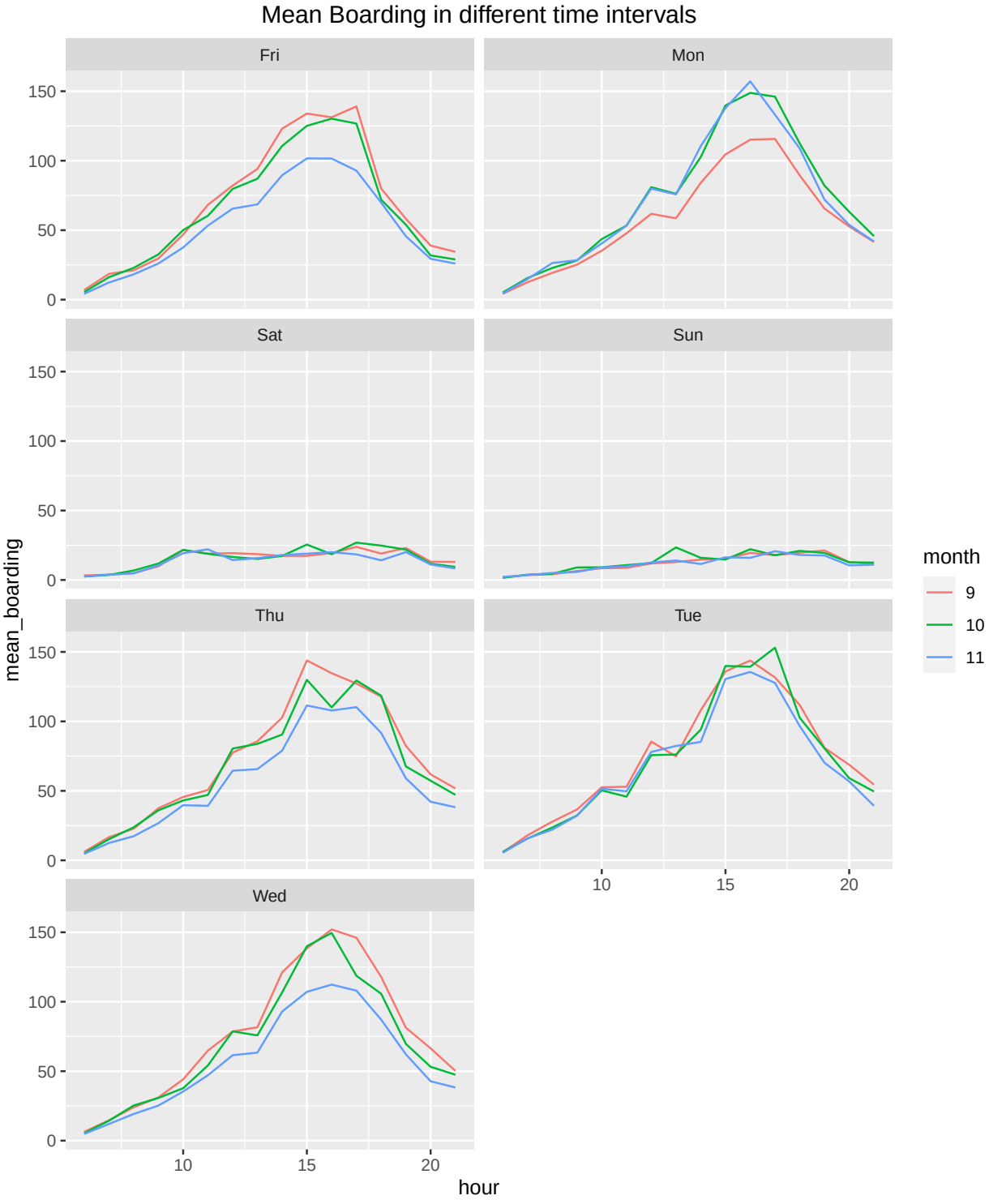
1 Problem 1: visualization

We first load capmetro_UT.csv and calculate average boardings grouped by hour,day_of_week and month.

```
## # A tibble: 5 x 4
## # Groups:   hour, day_of_week [2]
##    hour day_of_week month mean_boarding
##   <dbl> <chr>       <dbl>         <dbl>
## 1     6 Fri             9          6.88
## 2     6 Fri            10          5.44
## 3     6 Fri            11          4.2 
## 4     6 Mon             9          4.19
## 5     6 Mon            10          5.15
```

1.1 line graphs

(Q1_1_1) One panel of line graphs that plots average boardings grouped by hour of the day, day of week, and month. You should facet by day of week. Each facet should include three lines, one for each month, colored differently and with colors labeled with a legend. Give the figure an informative caption in which you explain what is shown in the figure



(Q1_1_2) Does the hour of peak boardings change from day to day, or is it broadly similar across days?

Based on the figure, It is broadly similar across days. At about hour 15 or 16.

(Q1_1_3) Why do you think average boardings on Mondays in September look lower, compared to other days and months?

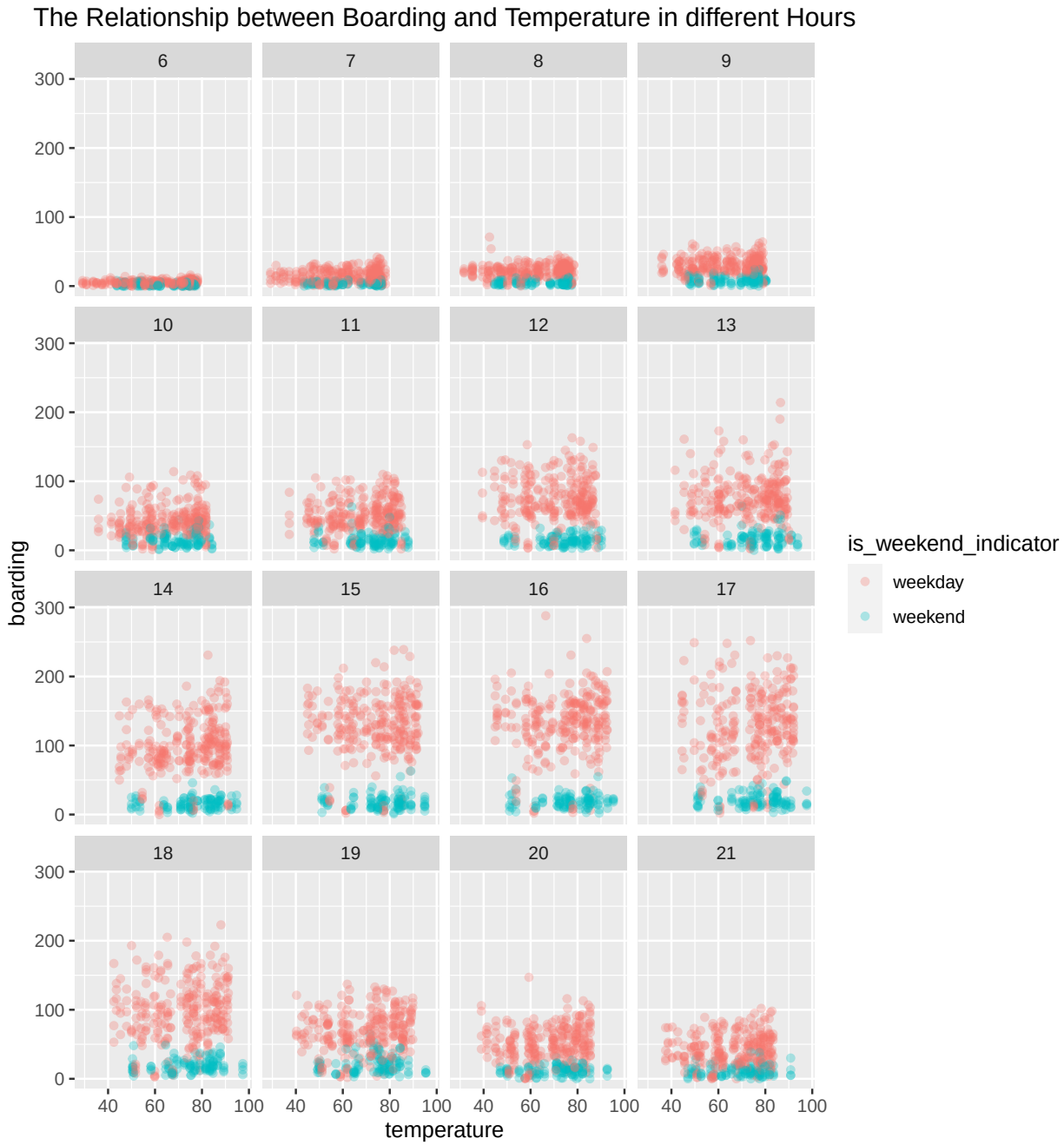
September is the beginning of Fall semester, which means there is less people on campus. On Monday, perhaps there is less courses than other days.

(Q1_1_4) Similarly, why do you think average boardings on Weds/Thurs/Fri in November look lower?

There are many midterm exams in November, which means students stay in the dorm to study for the exams without having to go outside.

1.2 scatter plots

(Q1_2_1) 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. Give the figure an informative caption in which you explain what is shown in the figure.



(Q1_2_2) When we hold hour of day and weekend status constant, does temperature seem to have a noticeable effect on the number of UT students riding the bus?

Just from the plot above, temperature doesn't have a noticeable effect on the number of UT students riding the bus.

2 Problem 2: Saratoga house prices

2.1 The Best Linear Model

(Q2_1) Build the best linear model for price that you can. It should clearly outperform the "medium" model that we considered in class. Use any combination of transformations, engineering features, polynomial terms, and interactions that you want; and use any strategy for selecting the model that you want.

The average of 5-fold corss-validation Rmse is used to evaluate a certain model. The cross-validation Rmse of middle model is 65989.29. Our target is very simple, to find a model with cross-validation rmse lower than 65989.29. A greedy algorithm is used for feature selection, and the result is as follows.

The best variables are(with cross-validation error of 57828.05):

```
# "price~livingArea+landValue+bathrooms+waterfront+newConstruction+
# heating+lotSize+centralAir+age+rooms+bedrooms+fuel+pctCollege+sewer+fireplaces"
# Error:57828.05
```

Corresponding cross-validation error is 57828.05, lower than 65989.29 from the medium model. The summary is as follows.

```
##
## Call:
## lm(formula = as.formula(best_str), data = SaratogaHouses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -228655  -35225   -4929   27480  457325
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.098e+05  1.968e+04   5.581 2.77e-08 ***
## livingArea               6.996e+01  4.615e+00  15.159  < 2e-16 ***
## landValue                9.219e-01  4.757e-02  19.379  < 2e-16 ***
## bathrooms                2.311e+04  3.370e+03   6.859 9.66e-12 ***
## waterfrontNo            -1.202e+05  1.554e+04  -7.734 1.77e-14 ***
## newConstructionNo        4.544e+04  7.308e+03   6.219 6.29e-10 ***
## heatinghot water/steam -1.045e+04  4.190e+03  -2.495 0.012679 *
## heatingelectric        -8.245e+01  1.232e+04  -0.007 0.994661
## lotSize                  7.599e+03  2.241e+03   3.391 0.000713 ***
## centralAirNo            -9.953e+03  3.478e+03  -2.862 0.004266 **
## age                     -1.304e+02  5.839e+01  -2.234 0.025600 *
## rooms                    3.020e+03  9.619e+02   3.139 0.001722 **
## bedrooms                -7.835e+03  2.567e+03  -3.052 0.002309 **
## fuelelectric            -1.093e+04  1.213e+04  -0.901 0.367799
## fueloil                 -4.381e+03  5.015e+03  -0.874 0.382466
## pctCollege              -1.102e+02  1.515e+02  -0.727 0.467139
## sewerpublic/commercial -1.524e+03  3.667e+03  -0.416 0.677775
## sewernone               -4.845e+03  1.712e+04  -0.283 0.777239
## fireplaces               1.037e+03  2.986e+03   0.347 0.728504
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58260 on 1709 degrees of freedom
## Multiple R-squared:  0.6534, Adjusted R-squared:  0.6498
## F-statistic:   179 on 18 and 1709 DF,  p-value: < 2.2e-16
```

In all, we successfully overperform the medium model!

2.2 The Best KNN

(Q2_2) Now build the best K-nearest-neighbor regression model for price that you can. Note: you still need to choose which features should go into a KNN model, but you don't explicitly include interactions or polynomial terms. The method is sufficiently adaptable to find interactions and nonlinearities, if they are there. But do make sure to standardize your variables before applying KNN, or at least do something that accounts for the large differences in scale across the different variables here.

Package kknn is used, and we slightly modify the evaluation function. Greedy algorithm is adopted again to select the best feature combination, the results are as follows. The best variables and cross-validation error for KNN is as follows

```
# "price~livingArea+landValue+age+pctCollege+waterfront+newConstruction"
# 58061.7316651973
```

The summary of KNN is:

```
##
## Call:
## train.kknn(formula = as.formula(best_str), data = SaratogaHouses,     scale = TRUE)
##
## Type of response variable: continuous
## minimal mean absolute error: 37716.64
## Minimal mean squared error: 3242600106
## Best kernel: optimal
## Best k: 11
```

2.3 Analysis

(Q2_3) Which model seems to do better at achieving lower out-of-sample mean-squared error? Write a report on your findings as if you were describing your price-modeling strategies for a local taxing authority, who needs to form predicted market values for properties in order to know how much to tax them. Keep the main focus on the conclusions and model performance; any relevant technical details should be put in an appendix.

The best variables and cross-validation error for KNN is

```
# "price~livingArea+landValue+age+pctCollege+waterfront+newConstruction"
# 58061.7316651973
```

The best variables and cross-validation error for linear model is

```
# "price~livingArea+landValue+bathrooms+waterfront+newConstruction+heating+
# lotSize+centralAir+age+rooms+bedrooms+fuel+pctCollege+sewer+fireplaces"
# 57828.05
```

Although the cross-validation error is lower for linear model, I still believe knn is better, as it uses only 6 variables to achieve its lowest error.

Moreover, $\frac{58061.7-57828.05}{57828.05} = 0.00404$, not very much.

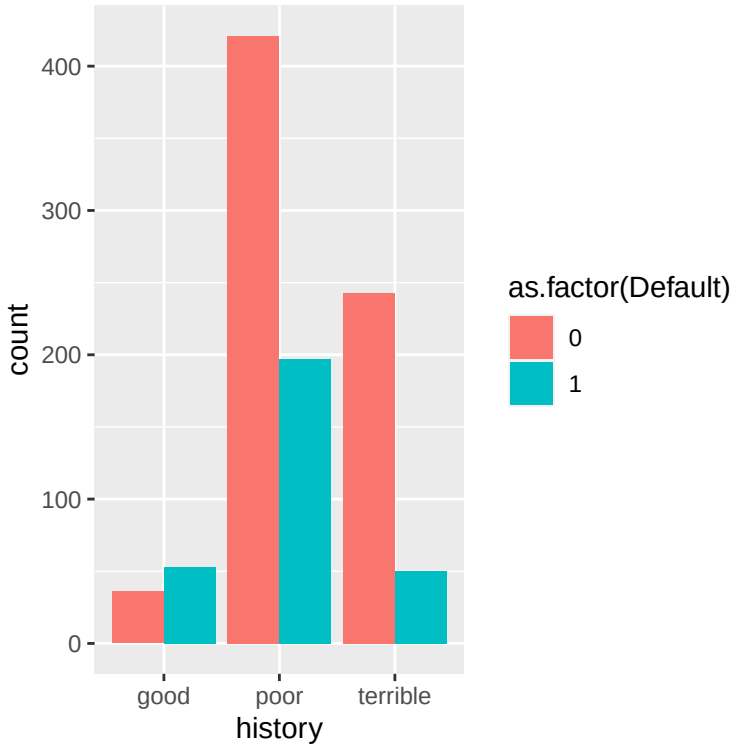
3 Problem 3: Classification and retrospective sampling

There are 300 Default and 700 not Default in the data.

As for history, good: 89 poor:618 terrible: 293.

(Q3_1) Make a bar plot of default probability by credit history

For categories poor and terrible we see the number of nodefault are greater than the number of default. For categories good we see the number of default is greater than the number of nodefault.



(Q3_2) Build a logistic regression model for predicting default probability, using the variables duration + amount + installment + age + history + purpose + foreign.

The summary of model is as follows.

```
##
## 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
```

(Q3_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?

What do you notice about the history variable vis-a-vis predicting defaults?

Based on the data, the default probability of people with good history is higher than the default probability of people with poor history. Not consistent with common sense! I don't think this variable work well in our model!(Although is significant)

What do you think is going on here?

There is a serious sampling problem. Due to "case-control" design, the default probability of people with good history is higher than other people. The sample can't reflect the real-world.

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?

This data set is NOT appropriate for building a predictive model of defaults. As the sample can't reflect the real-world.

Why or why not—and if not, would you recommend any changes to the bank's sampling scheme

I think the bank can just conduct random sampling of loans in the bank's overall portfolio. To solve unbalanced sample problem, the bank can change sample weight during modeling. Moreover, metrics like f1-score can be used to evaluate model performance.

4 Problem 4: Children and hotel reservations

4.1 Model building

(Q4_1)Using only the data in hotels.dev.csv, please compare the out-of-sample performance of the following models:

We first load the data. The mean f1-score from 5-fold cross-validation is applied to evaluate the performance of certain model.

1 Baseline model 1

The 5-fold cross-validation f1-score of Baseline 1:

```
## [1] 0
```

2 Baseline model 2

The 5-fold cross-validation f1-score of Baseline 2:

```
## [1] 0.4642258
```

3 Best linear model

We generate the time-stamp of year, month, day, and day of week from arrival_date.

Then, we use greedy algorithm to find the best feature combination.

The best feature combination is (cross-validation f1-score 0.50646)

```
## [1] "children~reserved_room_type+hotel+previous_cancellations+booking_changes"
```

The 5-fold cross-validation f1-score of best model:

```
## [1] 0.5064638
```

4 Analysis

The cross-validation f1-scores of 3 models are as follows:

Baseline1 model 1: 0

Baseline1 model 2: 0.4642258

Best model: 0.5064638

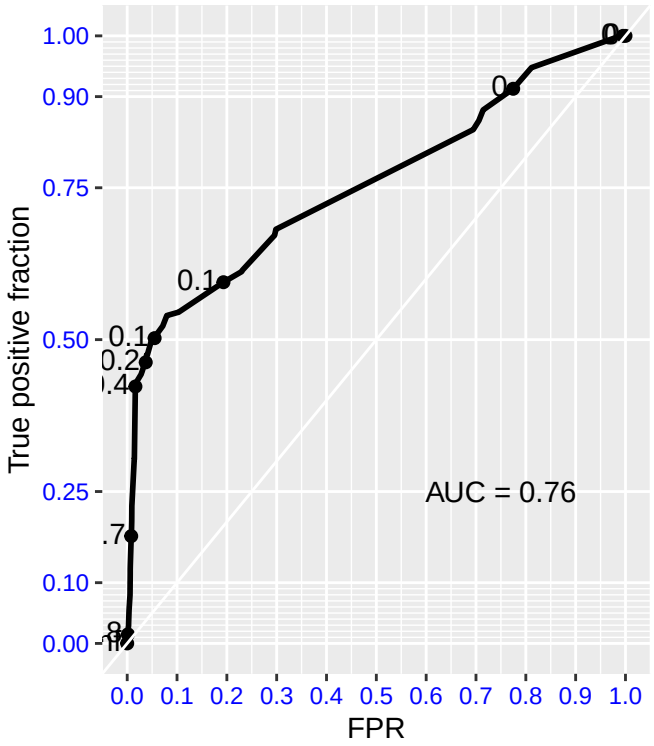
The best model has the highest f1-score.

4.2 Model validation: step 1

(Q4_2)Produce an ROC curve for your best model, using the data in hotels_val: that is, plot TPR(t) versus FPR(t) as you vary the classification threshold t.

We first fit on the hotelsdev data. Then we load hotels_val to conduct validation and draw ROC graph.

The plot is as follows:



4.3 Model validation: step 2

(Q4_3)How well does your model do at predicting the total number of bookings with children in a group of 250 bookings? Summarize this performance across all 20 folds of the val set in an appropriate figure or table.

We first fit on the hotelsdev data. Then we load hotels_val to calculate prediction accuracy.

The hotels_val is divided into 20 folds.

The following is the summary of expected number of bookings with children for that fold and actual number for that fold.

```
##    actual_num predict_num  difference
## 1          18    19.36443  1.36443344
## 2          20    22.02787  2.02787013
## 3          12    17.19036  5.19035783
## 4          17    16.96712 -0.03287646
## 5          23    19.25263 -3.74737039
## 6          25    23.38308 -1.61692053
## 7          28    21.43604 -6.56395728
## 8          17    17.04824  0.04823984
## 9          24    22.58401 -1.41599138
## 10         18    20.51186  2.51186271
## 11         14    21.29498  7.29497517
## 12         19    22.38645  3.38645423
## 13         17    21.76034  4.76033828
## 14         18    22.35551  4.35550900
## 15         19    18.82670 -0.17329682
## 16         20    18.23713 -1.76287476
## 17         24    22.78664 -1.21335581
## 18         24    23.60010 -0.39990154
## 19         21    21.28522  0.28522455
## 20         24    20.59058 -3.40942143
```

The following figure demonstrates the distribution of difference beween actual number and expected number among 20 folds.