HW 2

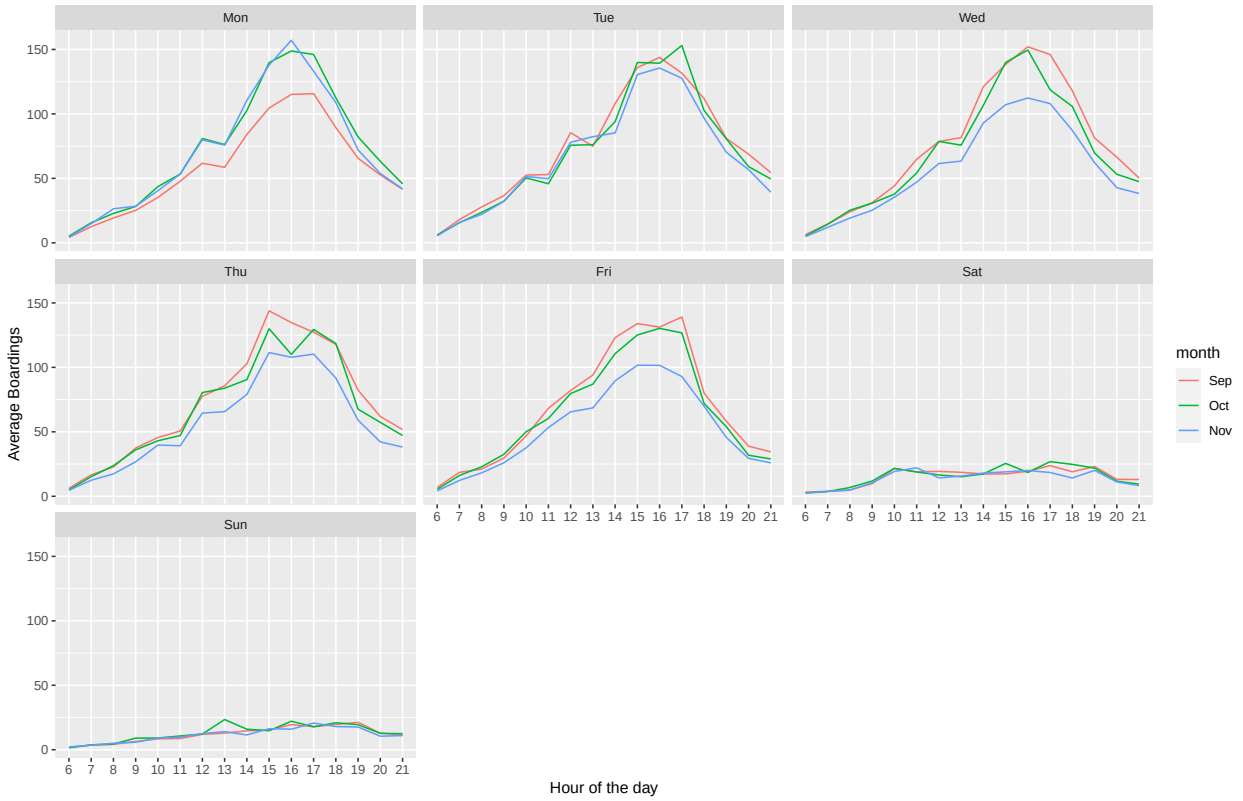
Ashesh Shrestha

3/9/2021

ECO 395: Exercises 2

Problem 1: Visualization

Figure 1.1 Average Boardings



As seen in the Figure 1.1, hour of peak boarding remains broadly similar across days. The highest average boarding is observed during 3 pm hour to 5 pm hour of each day. Another thing that we can observe in the figure is that average boarding on Mondays of September remains lower across hours of the day compared to other days and month. Similarly, we can also see that average boarding on Wednesdays, Thursdays and Fridays in November are lower. While it is difficult to accurately predict the reason behind so, one of possible reasons can be that the number of classes scheduled on Mondays during the month of September might be lower, whereas the number of classes scheduled on Wednesdays, Thursdays and Fridays could be lower on November.

Figure 1.2 Boardings v temperature in each 15−minute window faceted by hour of the day

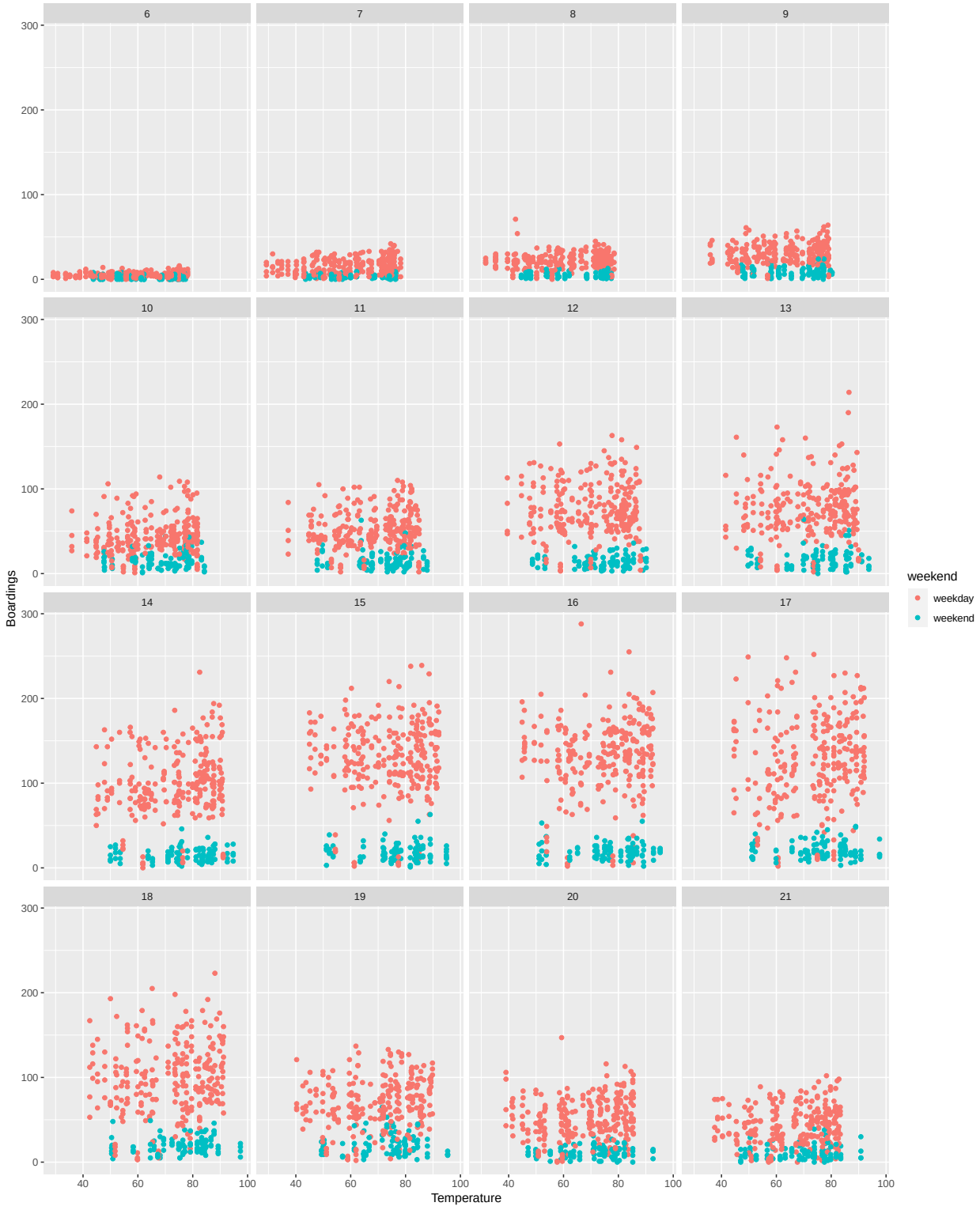


Figure 1.2 presents the scatterplot of boardings by temperature on weekdays and weekends faceted by hours of the day. As evident in the figure, when we hold hours of the day and weekend status constant, we do not see any noticeable effect of temperature on the number of UT students riding the bus.

Problem 2: Saratoga house prices

Mean RMSE for medium model

```
## [1] 65402.69
```

Mean RMSE for main linear model

```
## [1] 59772.66
```

Mean RMSE for k nearest neighbors with various k values

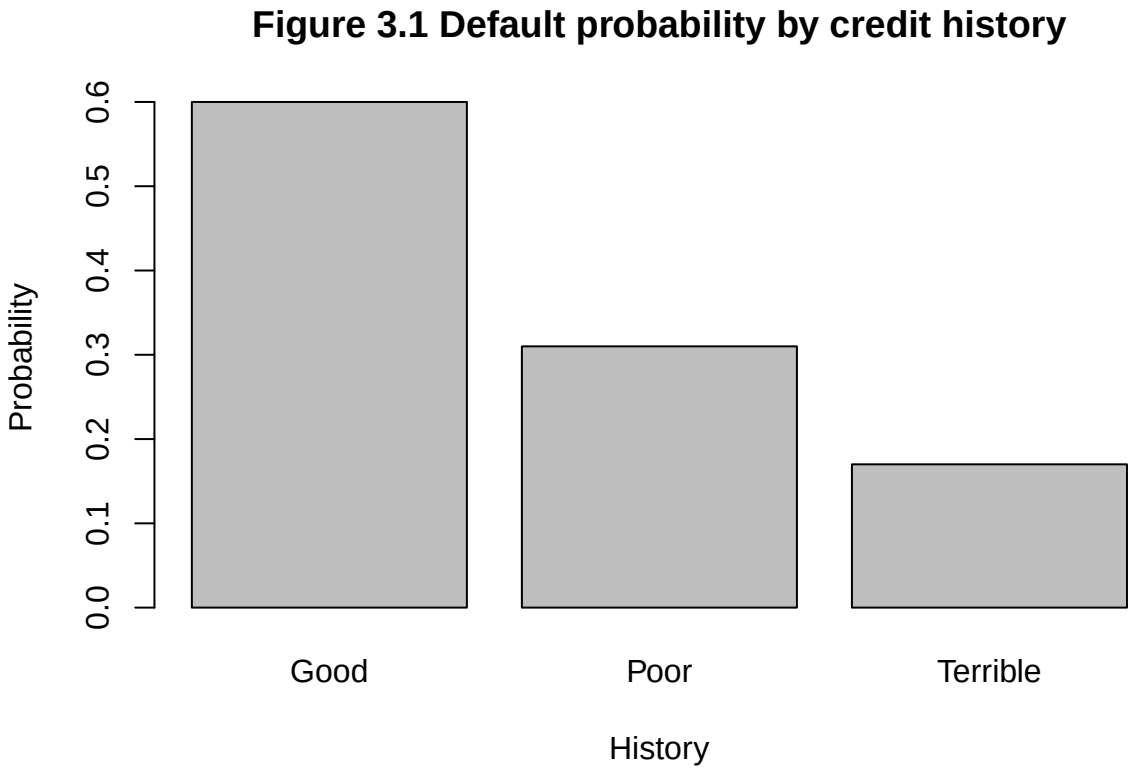
```
##              k     RMSE
## result.1     2 68918.48
## result.2     5 63732.13
## result.3    10 62516.70
## result.4    20 62455.85
## result.5    50 63650.38
## result.6    75 64481.68
## result.7   100 65298.82
## result.8   200 68003.53
## result.9   300 70541.11
## result.10  400 72339.79
```

I have built a linear and k-nearest neighbors regression models for prediction of house prices. I have chosen several features which affects the price of a house. For the linear model, I have included lot size, age, land value, living area, number of bedrooms, number of bathrooms, number of fireplaces, number of rooms, type of heating system, type of fuel used, availability of central air, interaction between age and central air, interaction between lot size and land value, interaction between number of bedrooms and rooms, and interaction between number fireplaces and type of heating system for predicting house price.

Similarly for the k-nearest neighbors, I have chosen lot size, age, land value, living area, number of bedrooms, fireplaces, number of bathrooms, number of rooms. type of heating system, type of fuel, and availability of central air as feature variables. As the k- nearest neighbors model is adaptable to find interactions and nonlinearities, I have not included interaction between the feature variables in the model.

After running both the models and measuring the out-of-sample performance by averaging the estimates of out-of-sample root mean squared error (RMSE) over many different train/test splits, I have found that average RMSE of k- nearest neighbors with k=20 is lower and hence perform relatively better.

Problem 3: Classification and retrospective sampling




```
##          (Intercept)             duration               amount         installment
##                -0.71                 0.03                 0.00                 0.22
##                  age          historypoor      historyterrible          purposeedu
##                -0.02                -1.11                -1.88                 0.72
## purposegoods/repair       purposenewcar       purposeusedcar       foreigngerman
##                 0.10                 0.85                -0.80                -1.26
```

Figure 3.1 shows bar plot of default probabilities for people with different credit histories, namely good, bad and terrible. The probabilities are quite counterintuitive. The probability of default for borrowers with good credit history is 0.6 while that for bad credit history is just above 0.3 and one for terrible history is 0.17.

Likewise, the coefficient of poor history and terrible history derived from the logit model are -1.11 and -1.88 respectively. This implies that having poor credit history multiplies the odds of default by approximately 0.33 while having terrible credit history multiplies the odds of default by approximately 0.15. Looking at the probabilities and coefficients of the logit model derived from given data, we see that the odds of default increases as credit history improves .

As defaults were rare, the bank sampled a set of loans that had defaulted for inclusion in the study. It then attempted to match each default with similar sets of loans that had not defaulted, including all reasonably close matches in the analysis. This resulted in a substantial oversampling of defaults, relative to a random sample of loans in the bank's overall portfolio. Therefore, this data set is not 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. For doing so, the bank should resort to random sampling which would prevent oversampling of defaults.

Problem 4: Children and hotel reservations

Baseline 1: Mean RMSE

```
## [1] 0.2682306
```

Baseline 2: Mean RMSE

```
## [1] 0.233225
```

Main linear model: Mean RMSE

```
## [1] 0.2314782
```

Figure 4.1: ROC curve

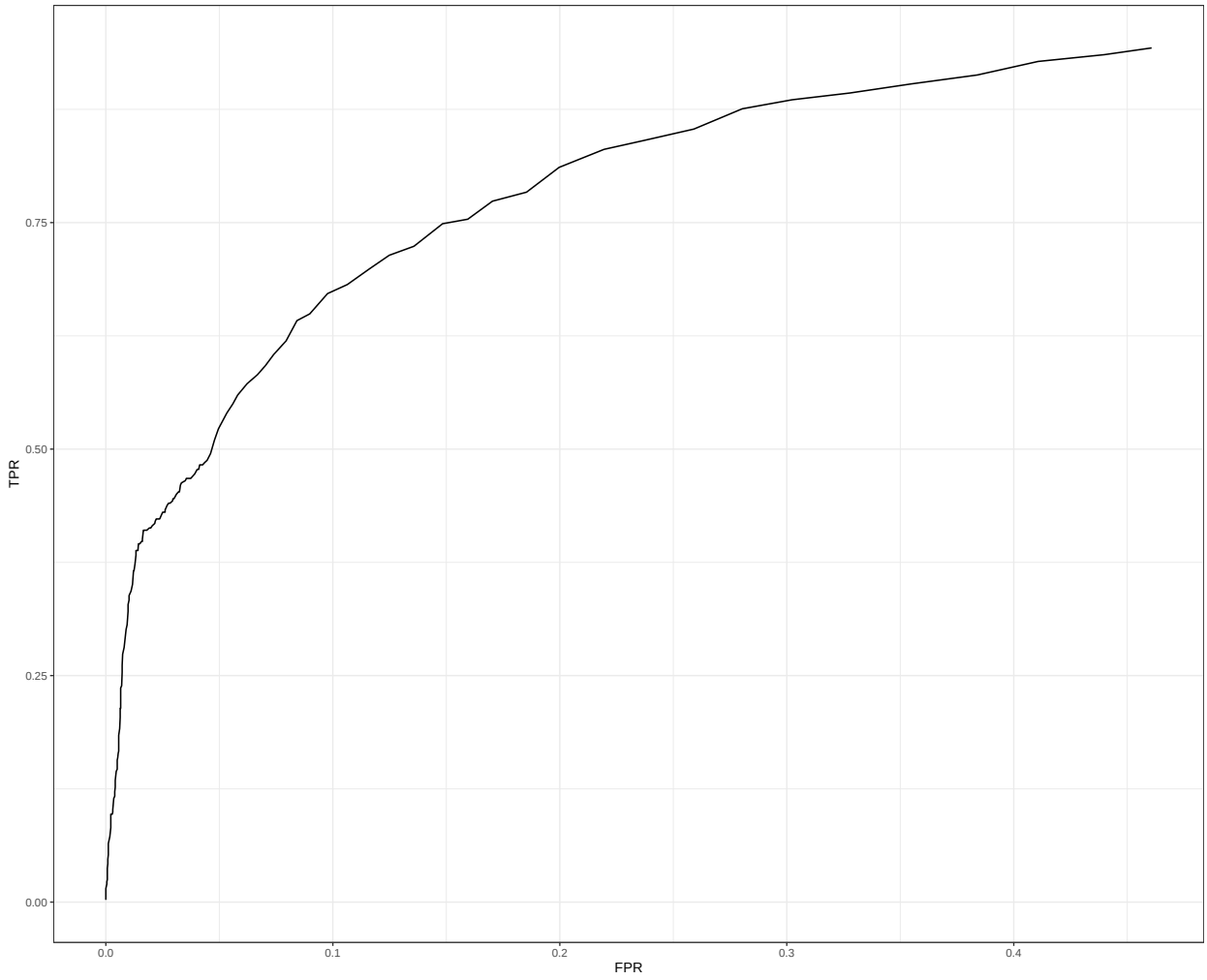


Table 1: Actual v expected number of bookings with childen

Fold_id	Actual	Expected
1	22	22.29369
2	21	19.67839
3	27	24.03965
4	22	21.47795
5	19	20.50929
6	18	22.02395
7	17	21.18100
8	19	20.79597
9	19	17.04718
10	24	20.30677
11	16	21.22584
12	13	17.77389
13	22	19.73028
14	28	20.84204
15	14	22.14672
16	22	21.95886
17	22	24.21831
18	20	20.98259
19	22	22.98225
20	15	21.67107

As seen in Table 1, out of 20 folds each with 250 observations, for 50 percent of the folds, the model has overpredicted the number of bookings with children. On the other hand, for 40 percent of the folds, the model has underpredicted the number of bookings with children. The model has exactly predicted the number of bookings with children for 10 percent of the folds. The maximum difference between actual and expected number of bookings with children in any fold is 8. Given the results, I think my model performed pretty well in predicting the total number of bookings with children.