ECO395M: PS2

Yuhui Gu and Zihao Yin

Problem 1

Table 1: Mean RMSE's of Two Models

Benchmark	New Model
66336.1	57912.39

With the goal of predicting house prices in Saratoga, we have constructed a linear model and a k-nearest neighbours (KNN) model based on the available dataset. Our linear model is an improvement to the model shown in class, and the performance of the models are measured by the mean out-of-sample root mean squared errors (RMSE) using 10-fold cross validation. Our improved linear model includes every independent variables present in the dataset and numerous interaction and polynomial terms. (More details of the model are shown in **Appendix A**). The mean out-of-sample RMSEs of the linear models — the benchmark model shown in class and our improved model — are displayed in Table 1.

Moreover, we have constructed a KNN model using standardised independent variables (more details of the model are shown in **Appendix A**), and the mean RMSE's using 10-fold cross validation at each level of k are plotted in Figure 1. Based on the figure, we find that the best performance occurs at a k of 9 with a mean RMSE of around 6.25×10^4.

Therefore, the improved linear model outperforms both the original benchmark and the KNN model. Although the linear model still retains a substantial error, it is the best model we have at the moment.

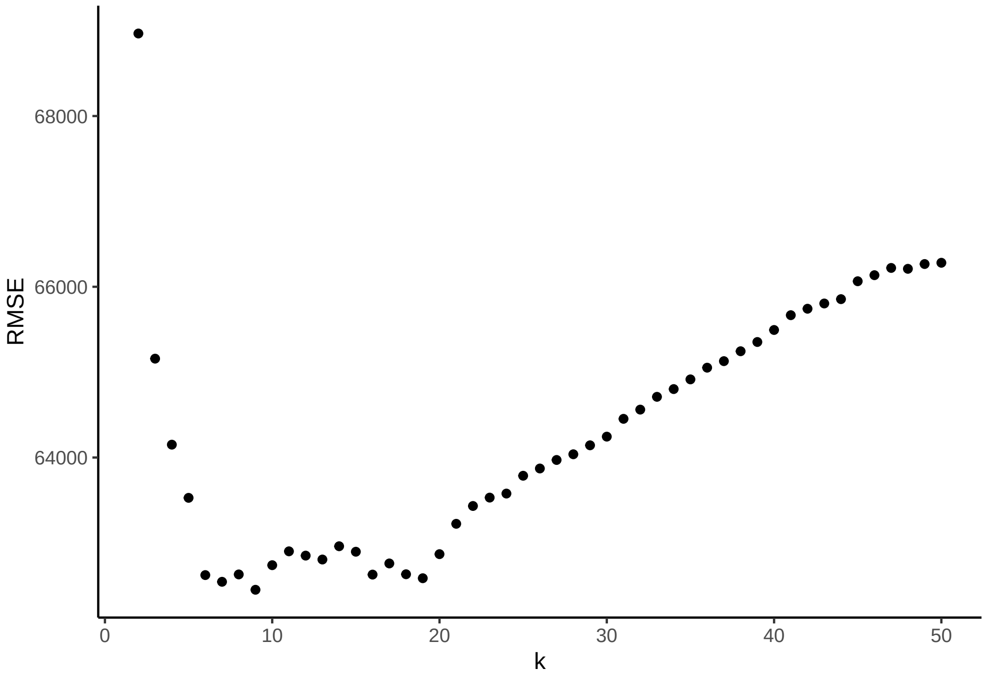

Figure 1: RMSE of KNN Regression

Problem 2

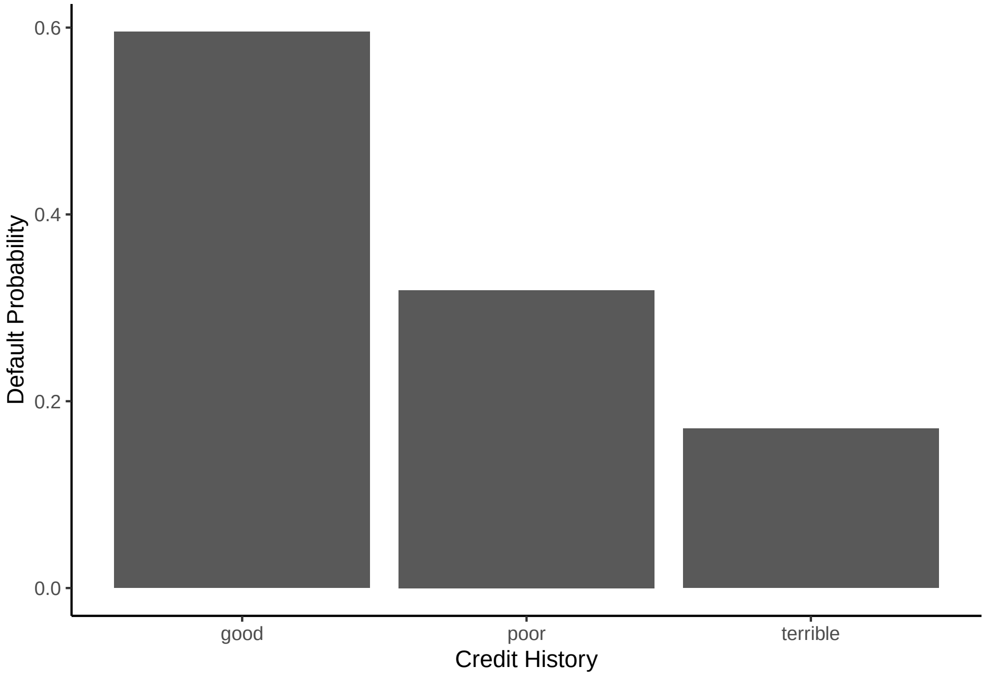

Figure 2: Probability of Default by Borrower's Credit History

The default probabilities for each credit history in our sample — good, poor, and terrible — are calculated as simple ratios of number of defaults to the total number of borrowers. The probabilities are plotted in Figure 2. We found that borrowers with better credit history have a higher likelihood of default. This is because there are relatively more borrowers with good history in the default subsample.

We then performed a logistic regression by controlling other characteristics of the borrowers, and the regression coefficients are shown in Table 2. Again, we find that borrowers with better credit history have a higher likelihood of default, as the coefficients of *historypoor* and *historyterrible* are both negative.

Table 2: Coefficients of Logit Model

	Dependent variable:
	Default
duration	0.025*** (0.008)
amount	0.0001*** (0.00004)
installment	0.222*** (0.076)
age	-0.020^{***} (0.007)
historypoor	-1.108^{***} (0.247)
historyterrible	-1.885^{***} (0.282)
purposeedu	0.725* (0.371)
purposegoods/repair	0.105 (0.257)
purposenewcar	0.854*** (0.277)
purposeusedcar	-0.796^{**} (0.360)
foreigngerman	-1.265^{**} (0.577)
Constant	-0.708 (0.473)
Observations	1,000
Log Likelihood	-534.977
Akaike Inf. Crit.	1,093.954
Note:	*p<0.1; **p<0.05; ***p<0.01

In a case-control study, there are "matching" variables and "exposure" variables. The matching variables should be similar across the default and non-default sets, and the exposure variable should not be controlled, so that the relation between the outcome and exposure variable can be uncovered. It is likely that the exposure variable is credit history in this study, while other features are used for matching. Given that there are very few borrowers with good credit history in the non-default set, it is likely that the borrowers with good credit history in the default set are similar in characteristics, except in credit history itself, to their bad history counterparts, meaning that they are actually more likely to default, and they were granted a loan just on the basis of their good credit history. Therefore, the dataset has an inherent selection bias.

In light of our finding, this dataset is not appropriate for building a predictive model of default because of the bias inherent in the sample. We would recommend the bank use a larger dataset that more accurately captures the characteristics of all borrowers to build predictive models.

Problem 3

Table 3: Mean RMSE's of Three Models

Benchmark 1	Benchmark 2	Our Model
0.268	0.233	0.223

Here we compare out-of-sample performances of three different models using 10-fold cross validation, like we did in Problem 1. The RMSE's of the models are shown in Table 3.

Our linear model includes every covariates that are present in the second baseline model. The *arrival_date* variable is instead split into *year*, *month*, and *day*, and all three are included in our model, with *year* and *month* as categorical variables. Additionally, each level in *adults* is included as a dummy variable in the model. We also included numerous interaction terms among the covariates, and more details on this are shown in **Appendix B**. Despite the drastic increase in complexity, the RMSE of this model is only 4.5% better than the second benchmark model, much to our dismay.

Now, we will validate our model using a new dataset. We predict whether each booking has children using a threshold $t \in (0, 1)$: if the predicted value of the model is less than t, we classify the booking as having no children, and vice versa. At each threshold, we construct a confusion matrix of our classification result, with which we calculate the true positive rate (TPR) and the false positive rate (FPR). The ROC curve of our model is shown below (Figure 3); the top-right region of the curve corresponds to a low t, while the bottom-left corresponds to a high t.

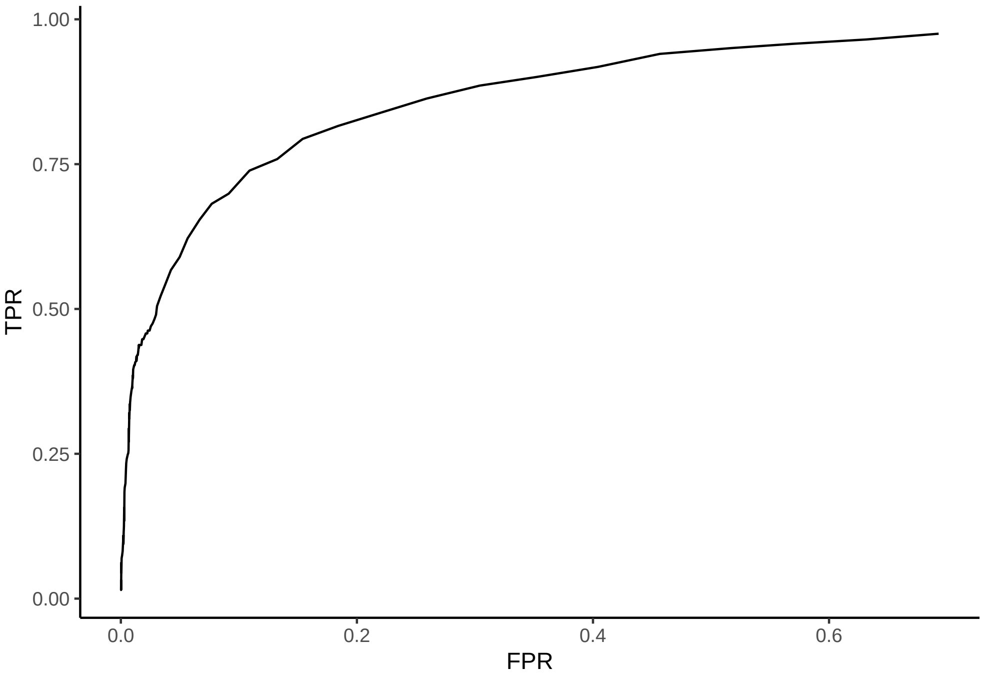

Figure 3: ROC of Our Best Linear Model

Furthermore, we randomly divided the validation set into 20 folds. For each fold containing around 250 bookings, we predicted the probabilities that a booking has children, and summed up these probabilities to obtain an estimate for the total expected number of bookings with children within a fold. The actual number of bookings with children and our estimated number for each fold are plotted below as a scatter plot (Figure 4).

The scatter plot of a perfect prediction would have every point falling on the diagonal line. So, based on the figure, we can see our prediction is quite poor, with the slope of the regression line being 1.1. Furthermore, the correlation between the actual values and our predictions is 0.66.

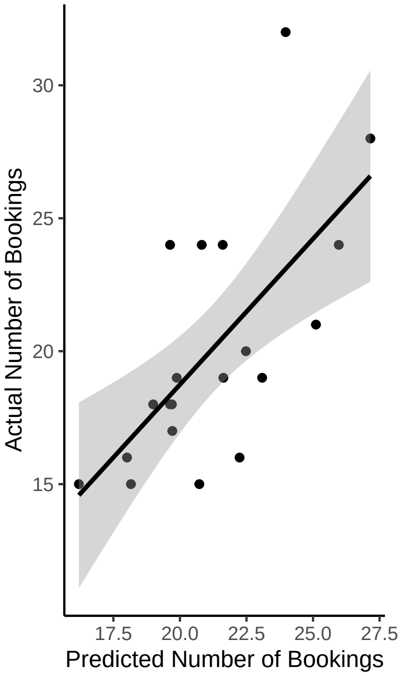

Figure 4: Performance of Our Linear Model

Appendix A

Linear Model

We have included every independent variable in the dataset in our linear model. We have also created a new regressor, *area.room*, that is the quotient of total living area to the number of rooms of a house. Other regressors that we have included are shown in the equation below, where *everything* denotes every column except for *price* in the original dataset:

$$\begin{aligned} price = {} & \beta_0 + \delta \cdot everything + \beta_1 age^2 + \beta_2 age^3 + \beta_3 rooms^2 + \beta_4 rooms^3 \\ & + \gamma_1 fuel \times heating + \gamma_2 heating \times fireplaces + \gamma_3 fuel \times \gamma_4 centralAir \\ & + \gamma_5 lotSize \times age + \gamma_6 landValue \times newConstruction + u \end{aligned}$$

The reason behind the inclusion of most of these regressors are purely empirical.

KNN

In our KNN model, we have included every independent variables of the dataset and *area.room*. All categorical variables – $fuel$, *heating*, *sewer* – are replaced by appropriate new dummy variables. All variables, including the dummy variables, are then standardised to ensure comparable weightings among the variables.

Appendix B

Our model can be expressed as follows

$$\begin{aligned} children = {} & \beta_0 + \beta \cdot everything + \gamma_1 year \times month + \gamma_2 day \times month \\ & + \delta_1 adults \times total\_of\_special\_requests + \delta_2 adults \times reserved\_room\_type \\ & + \delta_3 adults \times meal + \delta_4 adults \times stays\_in\_weekend\_nights \\ & + \delta_5 adults \times stays\_in\_week\_nights \\ & + \delta_6 adults \times is\_repeated\_guest + \delta_7 adults \times average\_daily\_rate \\ & + \phi_1 reserved\_room\_type \times required\_car\_parking\_spaces \\ & + \phi_2 reserved\_room\_type \times meal \\ & + \phi_3 reserved\_room\_type \times stays\_in\_weekend\_nights \\ & + \phi_4 reserved\_room\_type \times stays\_in\_week\_nights \\ & + \phi_5 reserved\_room\_type \times is\_repeated\_guest \\ & + \phi_6 reserved\_room\_type \times average\_daily\_rate + u \end{aligned}$$

where *everything* includes all covariates in the second baseline model, *year*, *month*, and *day*, and with *adults* as dummy variables.