PS2

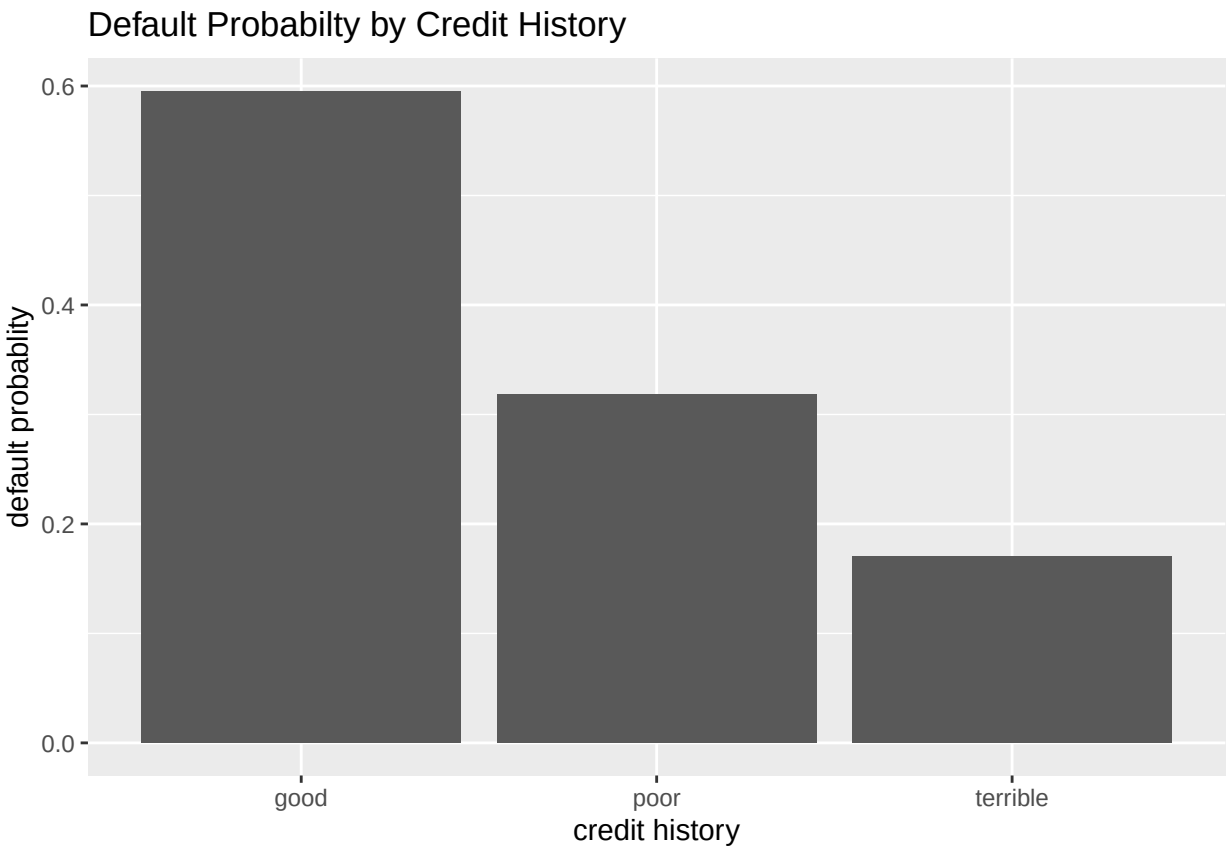
Chen-Yen Liu, Yu-Zhu Liu, Zi-yue Wang

1. The linear model which outperformed the medium linear model is: price = lotSize + age + landValue + livingArea + bedrooms + bathrooms + rooms + heating + waterfront + newConstruction + centralAir which was found using Stepwise regression.

To get the RMSE, I ran 20 times with randomly spilt samples for each model, we found that the linear medium model had an RMSE of 65627.990 and our chosen linear model had an RMSE of 58390.23. The KNN model had a RMSE of 67950.32 which was selected using repeated cross validation and then refit to the testing set. This means our chosen linear model was the best at predicting market values for properties in Saratoga. For a taxing authority it's clear that there are important factors in determining property value compared to the medium model: Land Value, Waterfront Property, and finally whether or not a house was a new construction.

2. People with poor credit history are much more likely to default than people with good and terrible history.

x axis is the credit history and y axis is the probability of defalt.



	x
(Intercept)	-2.73
X	0.00
checkingstatus1A12	-0.29
checkingstatus1A13	-1.77
checkingstatus1A14	-2.14
historypoor	-0.42
historyterrible	-1.85
purposeedu	0.28
purposegoods/repair	0.56
purposenewcar	0.98
purposeusedcar	-0.85
amount	0.00
savingsA62	-0.60
savingsA63	-0.48
savingsA64	-1.69
savingsA65	-2.08
employA72	2.57
employA73	1.98
employA74	1.11

	x
employA75	1.27
installment	0.33
statusA92	0.52
statusA93	0.02
statusA94	1.35
othersA102	-0.40
othersA103	-1.68
residence	0.08
propertyA122	-0.31
propertyA123	-0.12
propertyA124	1.13
age	0.02
otherplansA142	0.13
otherplansA143	-0.58
housingA152	-0.13
housingA153	-1.81
cards	0.51
jobA172	-1.95
jobA173	-0.77
jobA174	-0.31
liable	-0.02
teleA192	-0.85
foreigngerman	-0.37
rentTRUE	NA

History has a negative relationship with default, that means, people with poor and terrible credit history are less likely to be default. The result is consistent with what shown in the bar graph. This data set is not appropriate for building a predictive model of defaults, because it contains a high proportion of defaulted data, and reduces the model accuracy. My suggestion is to use a random data set that including similar proportion of default and non-default data.

Q3

###library and data

###Model Building

Initially, we will divide the data into training and testing sets, and proceed to develop baseline models 1 and Afterward, we aim to optimize the model by evaluating the p-value of each coefficient and investigating interaction terms.

Following the development of the optimal model, We hand-picked various features and interactions and eventually decided the above model as our final linear model, we construct a confusion matrix to compare its out-of-sample performance with that of other models. The confusion matrix showed a 0.9367 accuracy for out-of-sample prediction, which beat the second base line model by an absolute improvement of around 0.0001.

The accuracy scores of baseline1, baseline2, and the best model are provided below.

```
##    yhat
## y      0
##   0 8294
##   1  706
```

```
##    yhat
## y      0    1
##   0 8078  216
##   1  355  351
```

```
##    yhat
## y      0    1
##   0 8082  212
##   1  358  348
```

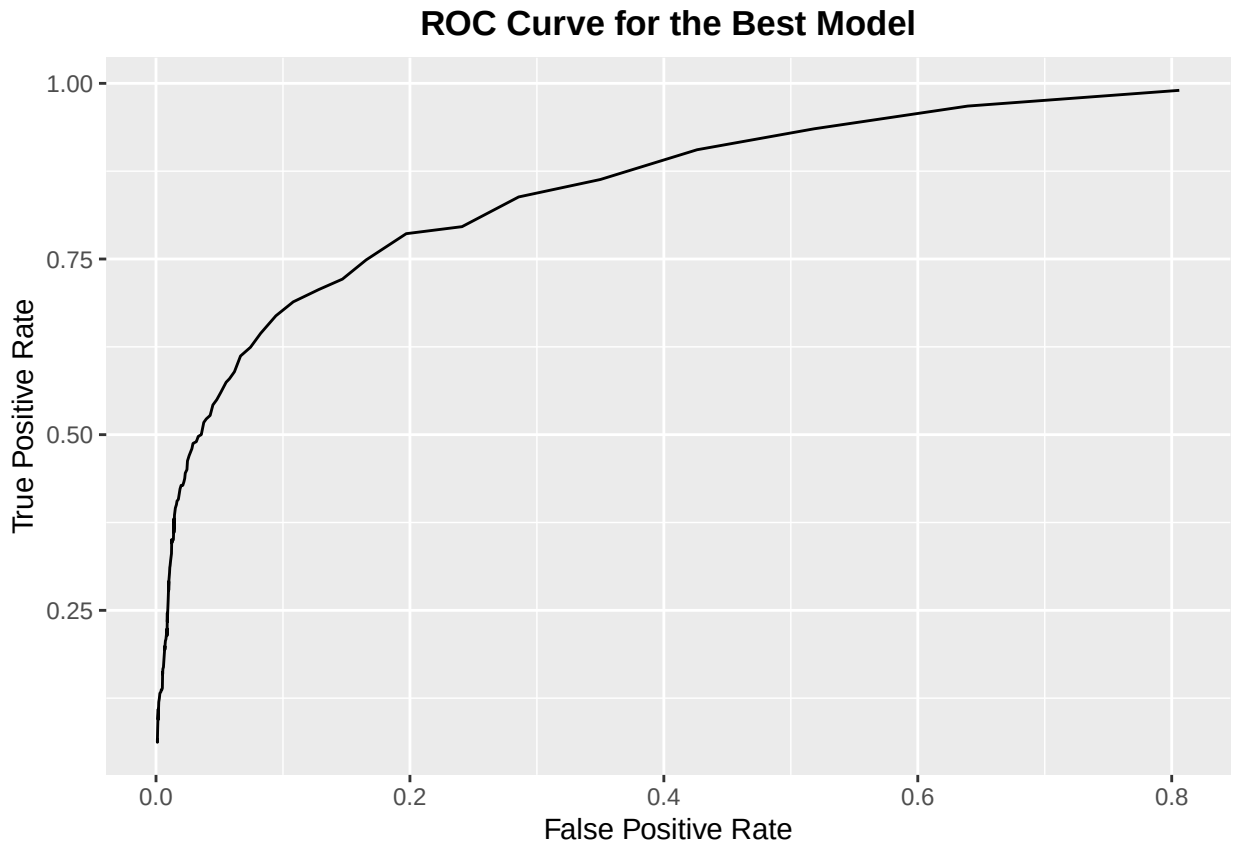
```
## [1] 92.16
```

```
## [1] 93.66
```

```
## [1] 93.67
```

Model Validation: Step 1

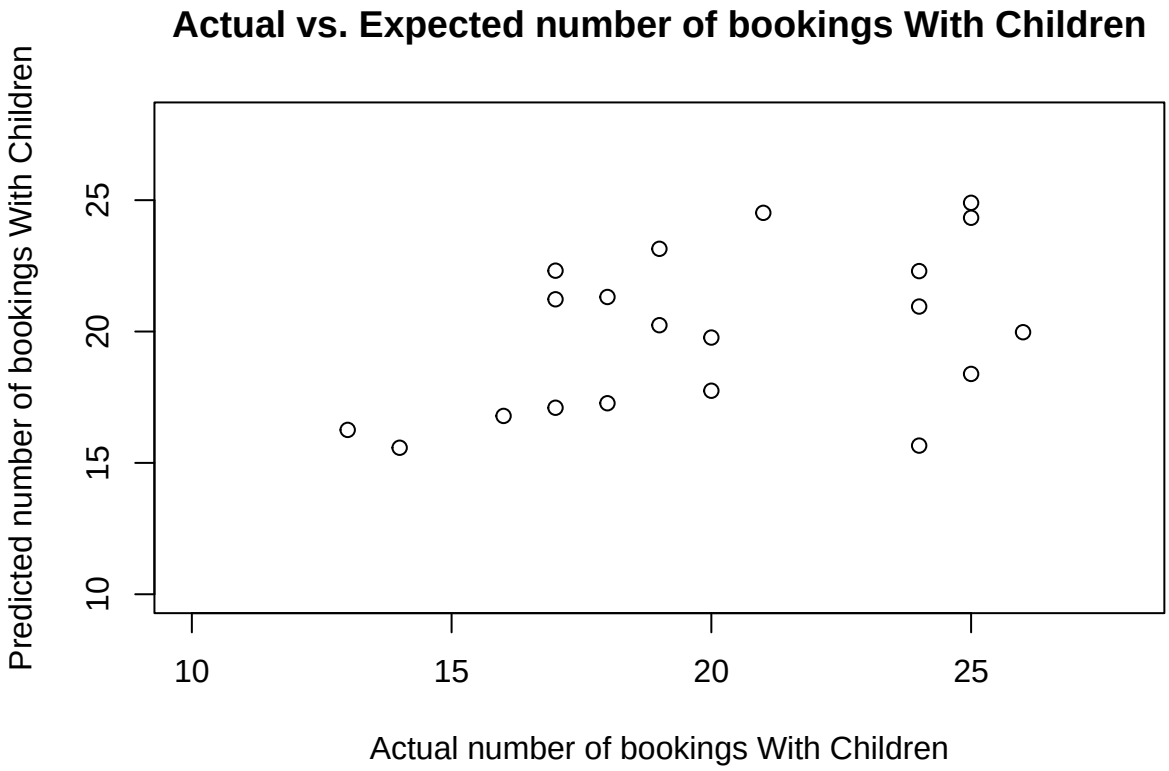
Validate our best model by testing on the `hotels_dev` data, and generate the ROC curve of this prediction using threshold of 0.01 to 0.9



Model Validation: Step 2

Perform 20-fold cross-validation on the hotels_dev dataset, where random fold numbers from 1 to 20 are assigned to each data entry using sampling.

For each fold, record the sum of predicted bookings and actual bookings to evaluate the performance of the model.




```
## [1] 2.859844
```

We calculate the difference between actual number of bookings with children and predicted number of bookings with children of each fold, and the mean of these differences is 2.86. We can see the expected numbers of bookings is only loosely following the actual numbers.