Exercise 2

Jinming Li Fan Ye Xiangmeng Qin

2/28/2024

ECO395M Homework 2

Jinming Li Fan Ye Xiangmeng Qin

2/28/2024

Problem1 Saratoga house prices

```
## [1] "Linear Model RMSE: 59201.5684264189"
```

```
##  [1] "KNN Model RMSE: 300.024081094876" "KNN Model RMSE: 279.025030555963"
##  [3] "KNN Model RMSE: 274.503193374298" "KNN Model RMSE: 269.974267481548"
##  [5] "KNN Model RMSE: 268.080079018104" "KNN Model RMSE: 266.10054930126"
##  [7] "KNN Model RMSE: 265.098271084645" "KNN Model RMSE: 264.743925733632"
##  [9] "KNN Model RMSE: 264.752314355396" "KNN Model RMSE: 263.284532754834"
## [11] "KNN Model RMSE: 262.903435368934" "KNN Model RMSE: 262.408253225279"
## [13] "KNN Model RMSE: 262.176874845124" "KNN Model RMSE: 261.846941715396"
## [15] "KNN Model RMSE: 261.606124951294" "KNN Model RMSE: 261.880599274728"
## [17] "KNN Model RMSE: 261.828320774977" "KNN Model RMSE: 261.76792114886"
## [19] "KNN Model RMSE: 261.888325197371" "KNN Model RMSE: 262.150025120321"
```

Report:

Result from linear model and KNN model

The initial class exercise provided three linear models with the following out-of-sample RMSE results: - **Linear Model lm1**: RMSE of 72,362.31 - **Linear Model lm2**: RMSE of 60,071.69 - **Linear Model lm3**: RMSE of 65,419.68

The new linear model we use is: price=β0 +β1×livingArea+β2×bathrooms+β3×age^2+β4×(livingArea×bathrooms)+…+ϵ, where "…" includes all the other variables present in the dataset as part of the initial full model. For the linear model, we applied feature engineering to create polynomial terms and interactions, notably a squared term for the age of the houses and an interaction term between living area and bathrooms. After ran 20 times with randomly spilt samples, our linearmedium model's RMSE is 57632.900565926 which is smaller than the RMSE of lm1, lm2, and lm3, indicating it outperforms the "medium" model that we considered in class. Subsequently, we developed a KNN model that incorporated standardized variables to address scale disparities. The performance of the KNN model was markedly superior, with an RMSE averaging approximately 263.5 across different runs. This stark difference indicates a substantial improvement over the linear models.

Conclusion

The KNN model consistently achieved a lower RMSE, indicating a more precise prediction of property values. The decrease in RMSE indicates that the KNN model's enhanced predictive capability and adaptability to non-linear relationships within the data. We advise to use KNN regression model for the determination of property market values within Saratoga County, because it is more accurate and thus more likely to result in more reliable property valuations for tax assessment purposes.

Appendix: Technical Details

The linear model incorporated advanced features, including age squared and an interaction term between living area and bathrooms, which were selected based on their potential impact on the house price. The KNN model's preprocessing included scaling of both continuous and categorical variables.

Problem2 Classification and retrospective sampling

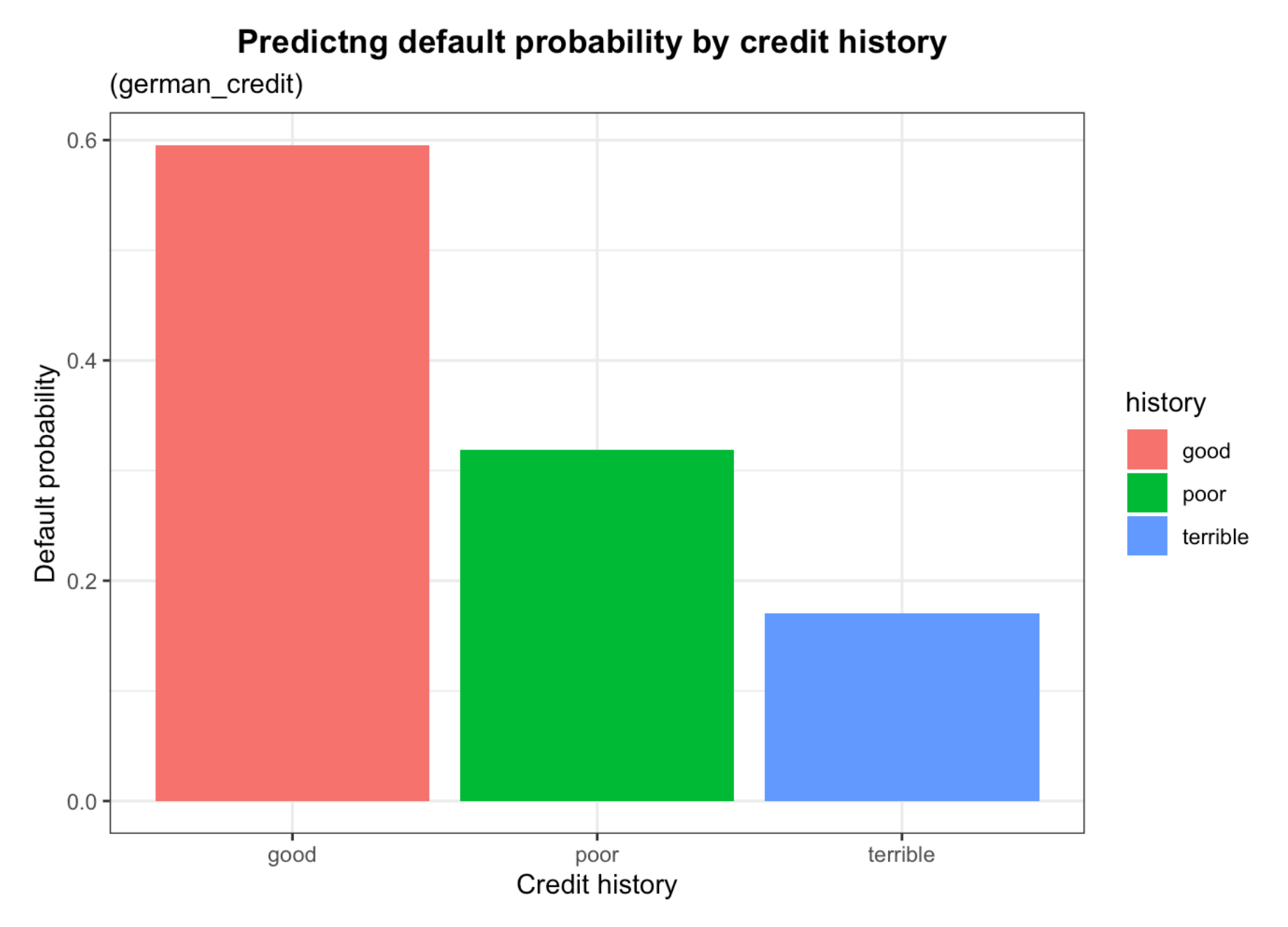


Figure 1:

Bar plot showing average default probability by credit history. We can see that among the three levels, loans with "Good" credit history have the highest default rate, while loans with "Terrible" credit history have the lowest default rate, indicating that a better credit history is associated with a higher loan default rate.

```
##         (Intercept)            duration              amount         installment2
##               -0.50                0.03                0.00                 0.24
##        installment3        installment4                 age          historypoor
##                0.54                0.64               -0.02                -1.45
##     historyterrible          purposeedu purposegoods/repair       purposenewcar
##               -2.07                0.73                0.40                 1.11
##      purposeusedcar       foreigngerman
##               -0.45               -1.41
```

```
##    yhat
## y     0   1
##   0 129   9
##   1  54   8
```

```
## [1] "Number of 'good'=  89"
```

```
## [1] "Number of 'poor'   =  618"
```

```
## [1] "Number of 'terrible'=  293"
```

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

We can observe a significant disparity between category counts, suggesting that the oversampling of certain specific categories within the data could be a potential reason for counter-intuitive statistical results. Loans marked with "good" credit history are underrepresented in the dataset, and a large portion of them are defaulted loans.

If the bank primarily focuses on borrowers with poor and very poor credit ratings when selecting samples, then the predictive model is likely to be biased towards predicting a high probability of default. Such bias means that the model's predictions may not accurately reflect the actual situation because it does not take into account a broader and more diverse range of borrower types.

To improve the accuracy of the predictive model, it is advisable for the bank to use random sampling methods to collect data, or at least to use a much larger sample size. This would include more borrowers with good and fair credit, providing a more balanced dataset to train the predictive model and better estimate the probability of default.

Do you think this data set is appropriate for building a predictive model of defaults

NO

138/200=0.69

A dissatisfying accuracy rate of 69% may indicate reasonable relationships between certain characteristics, but there seems to be elements within the data inhibiting successful predictions, as the intuitive response to predicting whether someone is likely to default (i.e., fail to repay a loan) does not align with the analytical outcomes.

The current dataset is unsuitable for predicting the "high" or "low" probability of borrower defaults due to biased sample selection. If the bank focuses on selecting samples mainly from borrowers with poor and very poor credit ratings, the predictive model is biased towards forecasting a high probability of default. This bias implies that the model's predictions may not accurately reflect reality as it fails to consider a wider and more diverse type of borrower. Additionally, the vast disparity in category counts suggests that oversampling of specific categories in the data could be a potential reason for the counter-intuitive statistical results.

To improve the accuracy of the predictive model, it is recommended that the bank employs a method of random sample selection when collecting data, or at least utilizes a much larger sample size. This approach would include more borrowers with good and average credit, providing a more balanced dataset to train the predictive model, thus enabling it to better estimate the probability of defaults.

Problem3 Children and hotel reservations

Model Building

First, we prepare for the data and build the baseline model.

We will then calculate the confusion matrix to look at our out-of-sample performance. Normally, we should choose t=0.5 as the predicted probabilties. But after several test, we consider it would be better for the hotels to place a high priority on not missing any bookings that might have children. As a result, we choose t=0.4 here.

Our linear probability model had the out-of-sample accuracy rate as a percentage:

```
##    yhat
## y      0
##   0 8240
##   1  760
```

```
##    yhat
## y      0    1
##   0 8120  120
##   1  440  320
```

```
##    yhat
## y      0    1
##   0 8112  128
##   1  440  320
```

```
## [1] 91.56
```

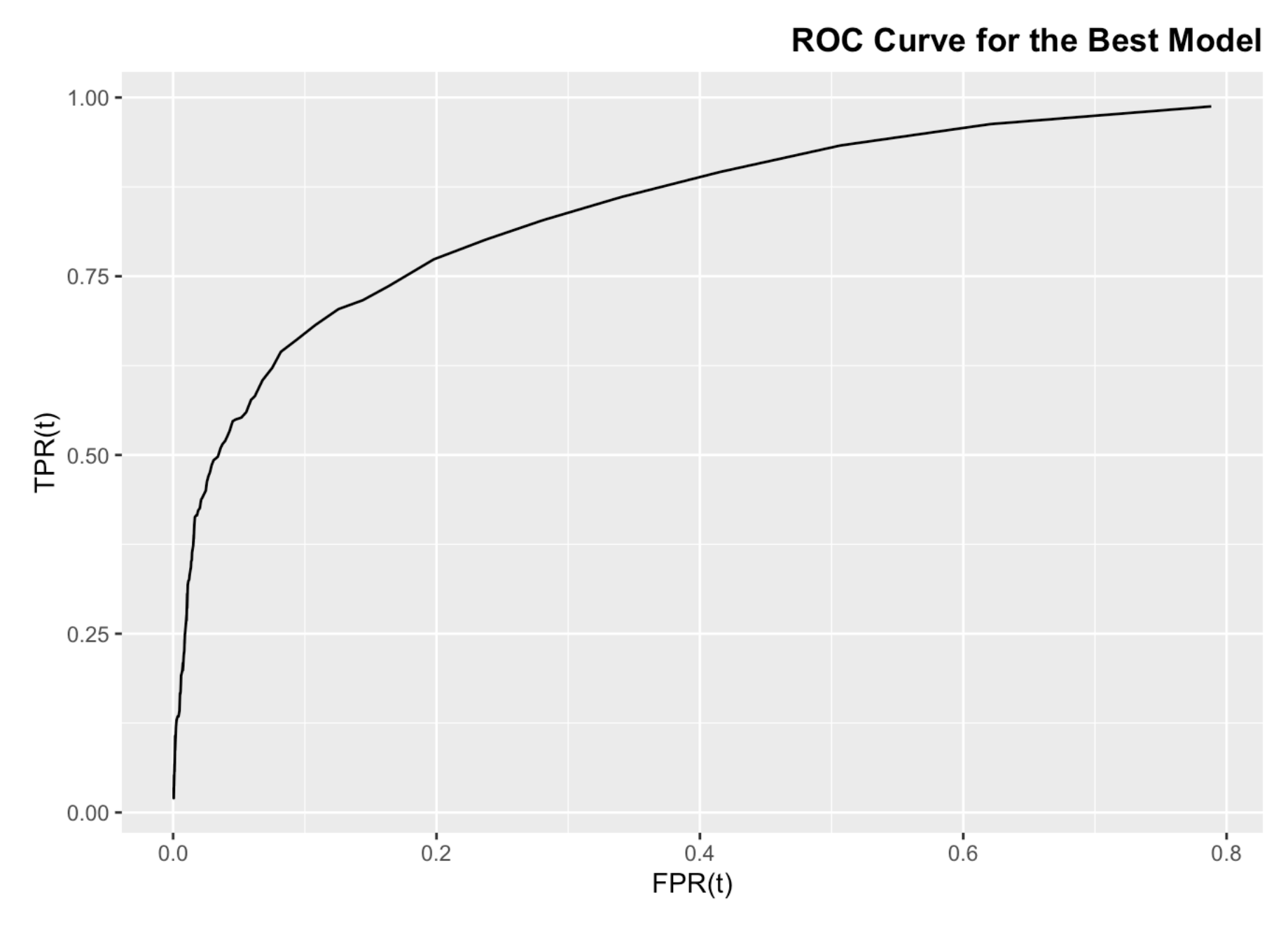
```
## [1] 93.78
```

```
## [1] 93.69
```

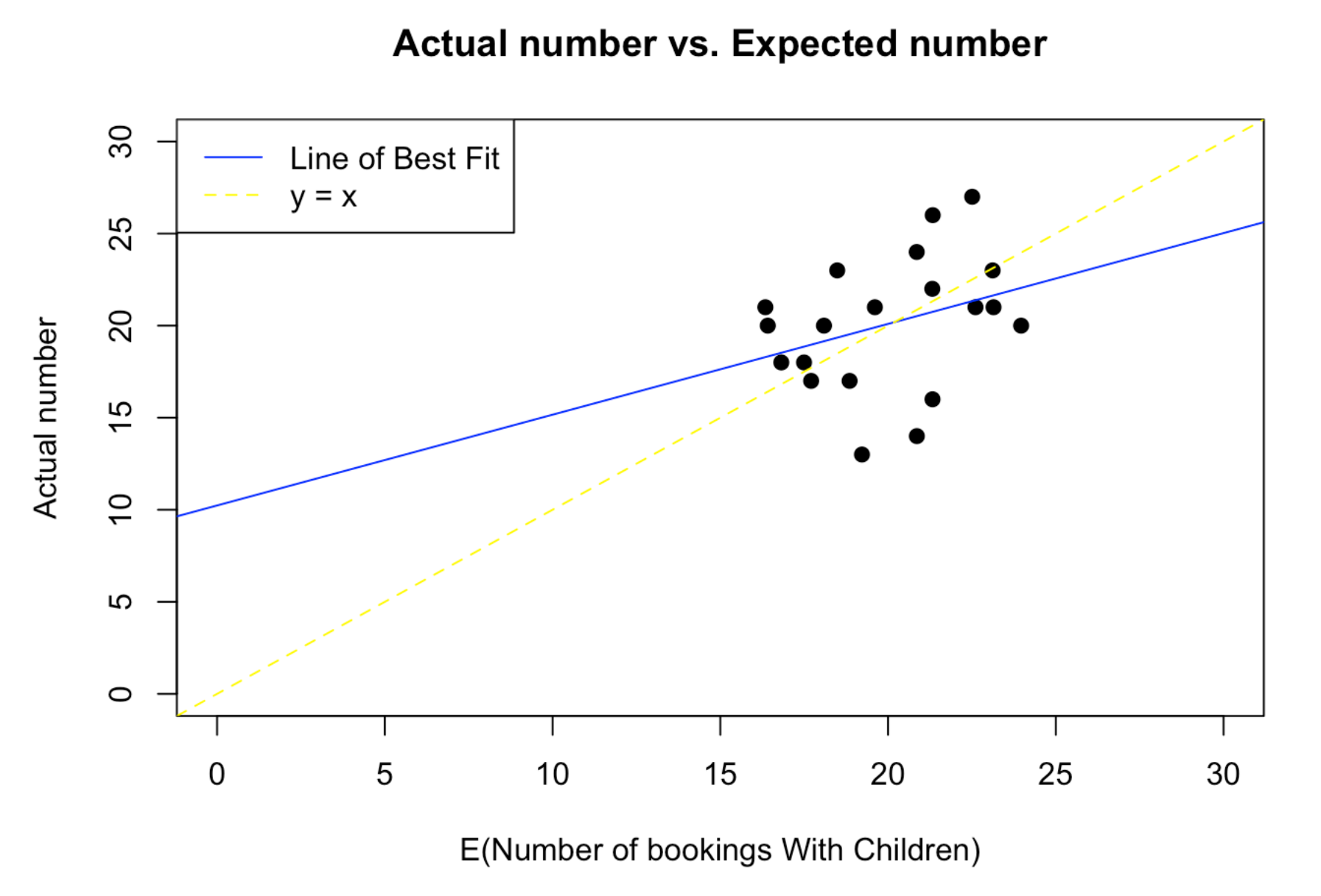
Model Validation

Validate the model from hotels_val, and generated a ROC curve with threshold of 0.01 to 0.95.

Showed as:



From the plot we can see that the optimal threshold is between 0.15-0.2




```
## [1] 2.976861
```

The difference between the expected number of bookings with children and the actual number of bookings with children is 2.9768608.

From the scatter plot, we can see that the line of best fit does not coincide with the identity line (y=x), which would represent perfect prediction. Instead, it is above the identity line. This indicates that the model tends to predict a higher number of bookings with children than actually occurred. In a word, the model's predictions can be described as approximate.

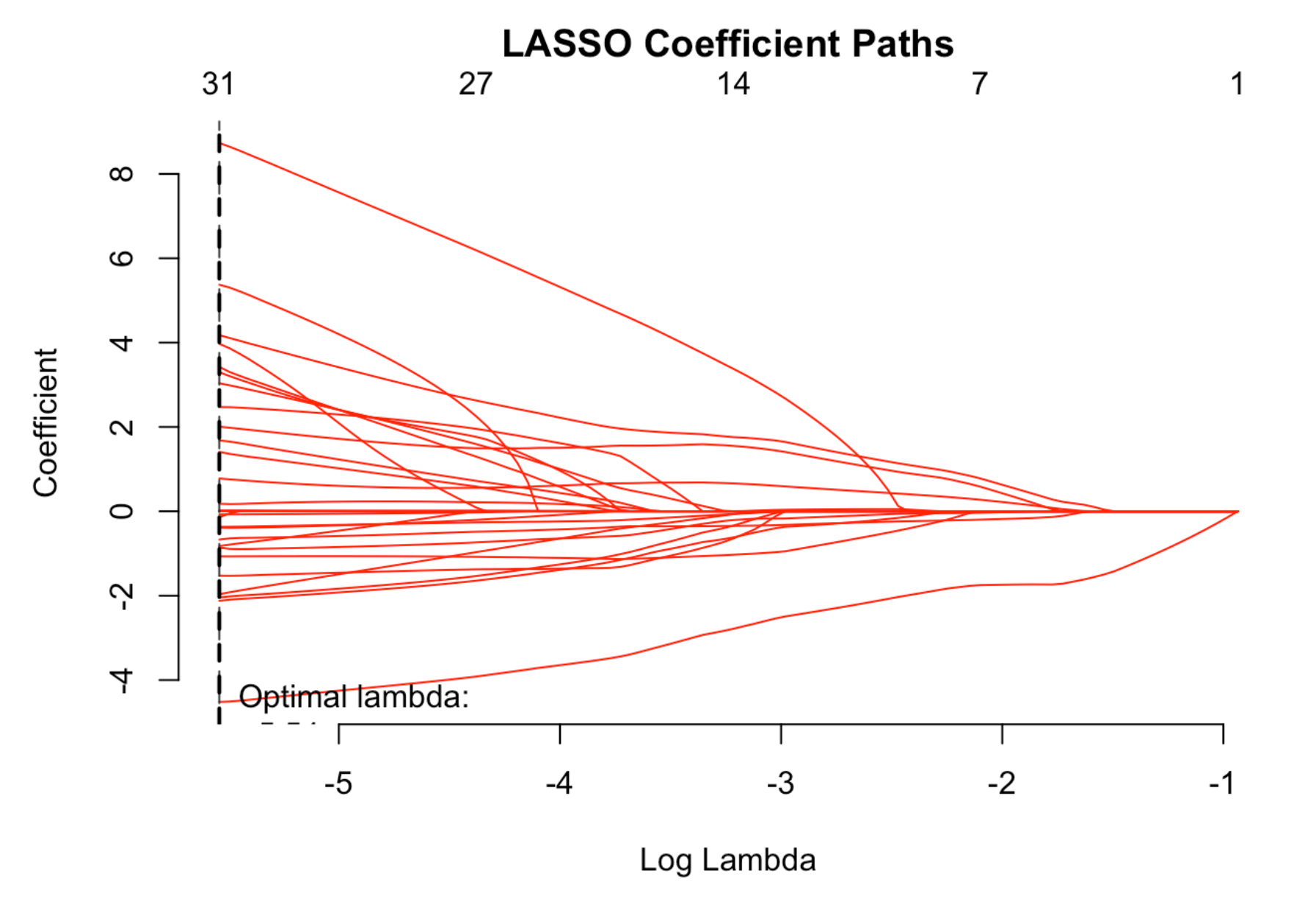
Problem4 Mushroom classification

Our goal is to build a model that can accurately predict the likelihood of a mushroom being poisonous based on the features provided. Since these features are categorical, they don't have a natural numerical representation that a mathematical model could process, thus we transform these categories using one-hot encoding.

We've removed the 'veil.type' feature before training our model. Because every mushroom has the same 'veil.type' value, meaning this feature doesn't vary. And then we employ lasso-penalized logistic regression for binary outcomes. After changing the category features into numbers and picking out the important ones with the lasso method, the generated plot is as followed:

```
## Levels for class : p e
## Levels for cap.shape : x b s f k c
## Levels for cap.surface : s y f g
## Levels for cap.color : n y w g e p b u c r
## Levels for bruises : t f
## Levels for odor : p a l n f c y s m
## Levels for gill.attachment : f a
## Levels for gill.spacing : c w
## Levels for gill.size : n b
## Levels for gill.color : k n g p w h u e b r y o
## Levels for stalk.shape : e t
## Levels for stalk.root : e c b r ?
## Levels for stalk.surface.above.ring : s f k y
## Levels for stalk.surface.below.ring : s f y k
## Levels for stalk.color.above.ring : w g p n b e o c y
## Levels for stalk.color.below.ring : w p g b n e y o c
## Levels for veil.type : p
## Levels for veil.color : w n o y
## Levels for ring.number : o t n
## Levels for ring.type : p e l f n
## Levels for spore.print.color : k n u h w r o y b
## Levels for population : s n a v y c
## Levels for habitat : u g m d p w l
```

```
##    seg100
## -5.540636
```

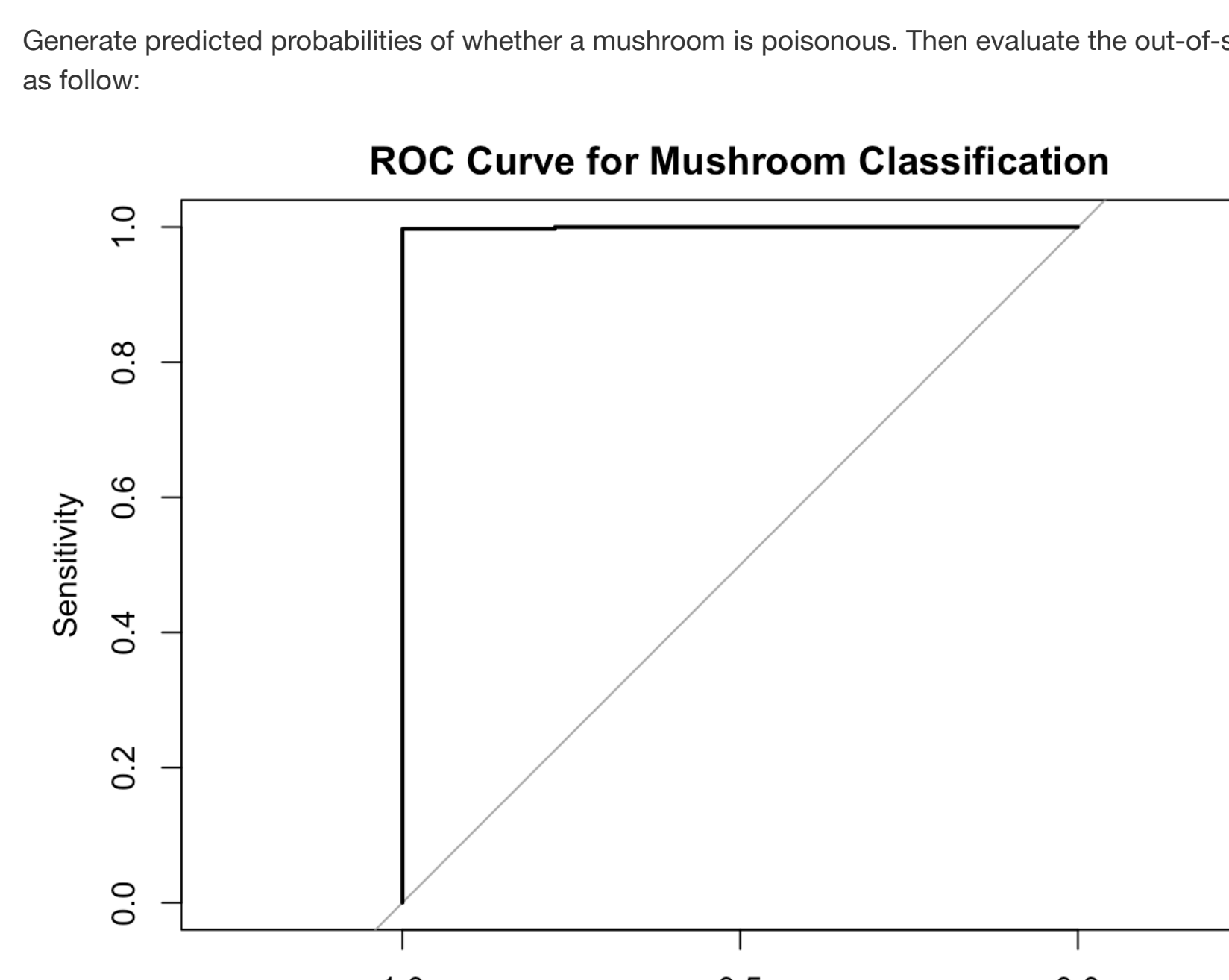



```
## integer(0)
```

Select non-zero variables:

```
##  [1] "intercept"                  "cap.shapec"
##  [3] "cap.surfaceg"               "bruisest"
##  [5] "odorc"                      "odorf"
##  [7] "odorl"                      "odorm"
##  [9] "odorn"                      "odorp"
## [11] "gill.spacingw"              "gill.sizen"
## [13] "stalk.rootc"                "stalk.rootr"
## [15] "stalk.surface.above.ringk" "stalk.surface.above.rings"
## [17] "stalk.color.above.ringc"   "stalk.color.above.ringy"
## [19] "stalk.color.below.ringc"   "stalk.color.below.ringy"
## [21] "veil.colory"                "ring.typef"
## [23] "ring.typen"                 "ring.typep"
## [25] "spore.print.colorn"         "spore.print.colorr"
## [27] "spore.print.coloru"         "spore.print.colorw"
## [29] "populationn"                "populationv"
## [31] "habitatw"
```

Generate predicted probabilities of whether a mushroom is poisonous. Then evaluate the out-of-sample performance by generating a ROC curve as follow:



The probability threshold for declaring a mushroom poisonous is:

```
## [1] 0.4166857
```

In the confusion matrix below, there are no cases where edible mushrooms were incorrectly predicted as poisonous (False Positives), as indicated by the zero in the e row and 1 column. There are 2 cases where poisonous mushrooms were incorrectly predicted as edible (False Negatives). The True Positive Rate (Sensitivity or Recall) is approximately 99.74%, indicating that nearly all poisonous mushrooms were correctly identified.

```
##    yhat
## y     0   1
##   e 864   0
##   p   2 759
```

```
## True Positive Rate (Sensitivity): 0.9973719
```

```
## False Positive Rate (1 - Specificity): 0
```

As shown above, the model demonstrates high accuracy and specificity. If the stakes are high, such as in health or safety applications, even a small number of false negatives may be significant, and further investigation into those cases would be warranted. However, the cost of misclassification, generalizability to unseen data, and the specific needs of its application still depends and may affect the value of this model.