Homework\_Thomas\_Kramer

1. What is the breakdown of employees who have excessive absences vs those who don’t?

*A: 36.6% of the employees have excessive absence*

![A screenshot of a cell phone

Description automatically generated]()

1. Fit a full logistic regression model to the data using the variables listed above. Based on this model, which variables do NOT appear to have a significant impact on the probability that an employee will have excessive absences in a month?

*A: Based on the results, it seems that the Body Mass Index, Pets and to lesser degree Social Smoking are not relevant.*

![A screenshot of a cell phone

Description automatically generated]()

1. What would be the next step in the modeling process?

*A: First assess the accuracy of the current model. Then next step re-run the model eliminating stepwise the least significant variables, while assessing the accuracy (misclassification rate etc.) at each step.*

*In my case (with all the 7 features) they initially look like this.*

*![A screenshot of a cell phone

Description automatically generated]()*

*![A screenshot of a cell phone

Description automatically generated]()*

1. Now suppose that you have iterated through the modeling process and have arrived at a logistic model that includes only Distance from Residence to Work, Service time, Son, and Social drinker as the explanatory variables. After fitting a model containing only these explanatory variables:

How does an employee’s years of service impact his or her likelihood to have excessive absences?

![A screenshot of a cell phone

Description automatically generated]()

*A: Years of Service have positive impact, recuing absence (negative coefficient, means it reduces the likelihood of absence.*

How does the number of children that an employee has affect his or her likelihood to have excessive absences?

*A: Children have negative impact, it increases the likelihood of absences (positive coefficient, means it increases likelihood of absence.*

1. Based on the confusion matrix, is the logistic regression model more successful at identifying employees WITH excessive absences, or those WITHOUT excessive absences? How can you tell?

![A screenshot of a cell phone

Description automatically generated]()

*A: The model is better at identifying the employees WITHOUT excessive absences*

*It does a fair job of identifying True Negatives (=employees WITHOUT Absence) It catches almost 65% of those correctly. It is however very poor at identifying the True Positives (=employees WITH absences), where it only gets 47% right.*

1. What is the misclassification rate for the model?

*A: Miss-classification Rate is 37.43 Percent.*

*This hides a bit the reality, of being poor. The majority (63.3%) of the employees are not absent during the month. If the classification would always have predicted the majority class (=not absent), it would have had a misclassification rate of 36.6%. So the classification we just did is very poor*

![A screenshot of a cell phone

Description automatically generated]()

**

1. What is the area under the ROC curve?

![A close up of a map

Description automatically generated]()

1. Based on all of this information, do you believe this final logistic regression model is useful for helping the company to identify absenteeism in employees? Why or why not?

*A: The model is rather poor. The Area and the Curve is only 0.63 and the Misclassification rate 37% is even worse than just choosing the majority class (not absent).*

*It seems very bad at identifying absenteeism after all. While in reality 271 employees misses more than a day, our model only classifies 96 persons to be absent (and gets more than half of them wrong!) Eliminating non-significant features actually increased the numbers of predicted absences but has many misclassified results.*

1. What recommendations might you make based on this model?

*A: I would be careful of using this model. It does not seem to make a good job at all. Maybe we are missing real important variables, which influence the absenteeism. The coefficients might give an indication that service time is (slightly positive) and children might be negative (more absence), but I would recommend to further investigate.*