**If you have partition your dataset and run your models please answer the following:**

1. **Have you reviewed the estimation results?**

Yes we have divided our data into validation and test. But before that we needed to clean our dataset. Fortunately, for us, we did not have any missing values and duplicates were out the question among thee outliers as well. The reason we have done that is because we want to assess the accuracy of our models in explaining the research questions we have asked. However, when we were doing it, we were facing the problem related to unbalanced data, since our variables are categorical. After having discussed in our meeting with you, we applied the ROSE library to balance the data out and then we moved forward with our data.

1. **Are the results the ones you expected? why yes or not?**

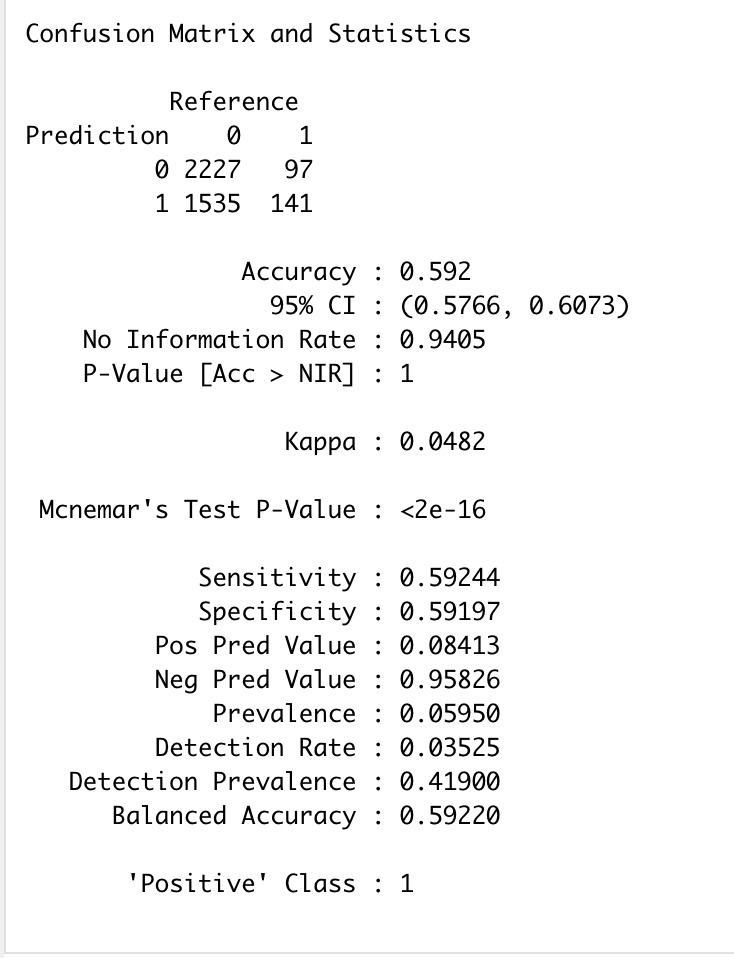
We were having some strange results before we went further, we took the median age by converting them into numerical and we did the same thing for income as well, both of which were numerical. After that the results were somewhat expected, depending on the model. It should be noted that we made 3 models. Model 1 is constructed by identifying the characteristics of customers buying caravan insurance.

1. **Do you need to run more estimations?**

As we saw that for both the models, we are not getting what we want. We tried addition and subtraction of several variables but to no avail. So in that case, we will be moving forward with a 3rd model based on our knowledge domain and try to fit it to our research questions.

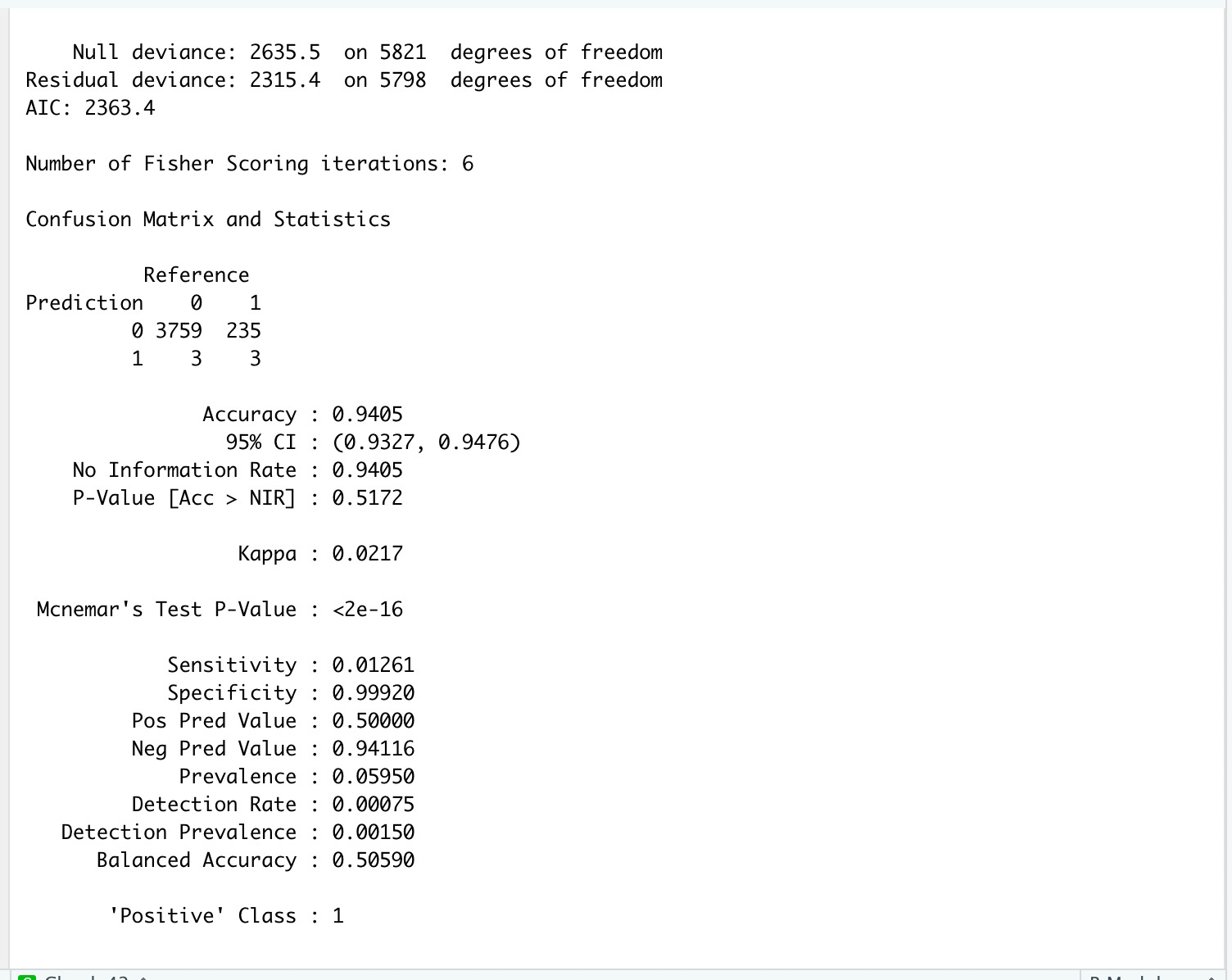
1. **Start drawing explanations for your results. Difference between models?**

Following is the result of model 1. As we can see that our accuracy is ~60% and Sensitivity is somewhat ~60% as well.



We then construct model II and compare it with model 1. For the construction of model II, we ran correlation matrix. We found that there were some variables above .75 of all the 86 variables. Naturally, we cannot take all of the variables and had to reduce them. So in order to do that we ran another correlation with caravan insurance. We excluded all the variables below .1 because of low correlation. In doing so, we reduced 86 variables to 58.

Next thing we did was that we applied GLM on those variables. The reason as to why we did that is because we want to find the probability of customer types and GLM model expresses it between 0 and 1. For this model, following are the results:



For the above model, our cut-off value is at .5. Our specificity is at 99% and accuracy is 94%. This may sound good but it is opposite of what our research question is. For that we need sensitivity to increase.

1. **At this stage, you should be assessing the predicting performance of your project.**

Yes, we have tried to predict and are now in the process of assessing the predictive performance of our models. Once, we have it, we shall report it in the next section.

1. **Show the indicators (error rates, accuracy rates, confusion matrix, decile-wise lift curves, lift curves) showing the predictability of your model**

We have already mentioned the results of our confusion matrix in Q2 and are now working on assessing the predictive power of our models. Since our models are categorical, we would use lift charts and decile wise charts to deduce something concrete about our model

1. **Explain your results**

We have drawn the confusion matrix for our 2 models for which a brief summary has been mentioned above. Right now, we are in the process of drafting model 3 which we assume would be able to give us our desired result. Once, we have that we will test the predictive power of our models and draw deductions from that. We look forward to mentioning that in the next update.