1. **Have you chosen your variables? Any challenges in data dimension?**

As our data set has 86 variables, 85 of those are categorical and 2 of them are numeric. Our first instinct was to apply PCA but much to our dismay we could not do it because our dataset was not numeric. We thought about going the longer route by converting variables of interest into binary and then applying PCA but the problem we encountered was that we would little room to explain our findings as the scope of binary was limiting us.

So then the challenge, was to select a distinct number of variables that could explain or shed some light on our research question, \*who is more like to buy insurance?\*. The problem we encountered was that not only did we have main categories of customers but rather sub-categories as well. So instead of taking all of those categories into account we grouped sub-categories into main categories.

In order to accomplish this we used the help of excel and pivot tables to determine which sub-category belonged to which main category. To our surprise, our grouping was surprisingly accurate. Hence, we now have 10 main categories which consists of their respective subcategories. This helps us with our 2nd research question, \*what are the characteristics of those likely to buy insurance?\*.

In order to find answers to these questions, we have identified the following variables of interest:

* Income
* Main customer type
* Sub Customer Type
* Caravan Insurance
* No. of houses
* No. of households
* Skilled
* Unskilled

1. **Have you chosen the technique (data mining task)?**

As mentioned above since our dataset is categorical, applying data mining techniques related to numerical data would be a total disaster. Hence, we have applied the basic rules of data mining which are as follows:

1. **Descriptive Analytics**
2. **ELT [Extract, Load, Transform] / Cleaning**
   1. Duplicate identification
   2. Missing values identification
   3. Visualization
      1. Bar Charts – because categorical data is best explained by it
      2. Histograms
      3. Boxplots
   4. Outlier identification
3. **Transformation**

In order to have a responsive model we have converted our variables of interest into binary. In addition to this, we have taken income and converted it into numeric form.

1. **Grouping**

Since we will be using Logit regression we had to group a few of our variables. We have 10 main customer types and 44 sub-customer types. The way we moved forward was that we computed sub customer types in to main customer types and then further grouped sub customer types into a new group. After doing this we have 10 sub customer types and 10 main customer types.

1. **Model Type**

We have used logistic regression analysis to uncover hidden patterns in relation to our research questions in order to make our prediction.

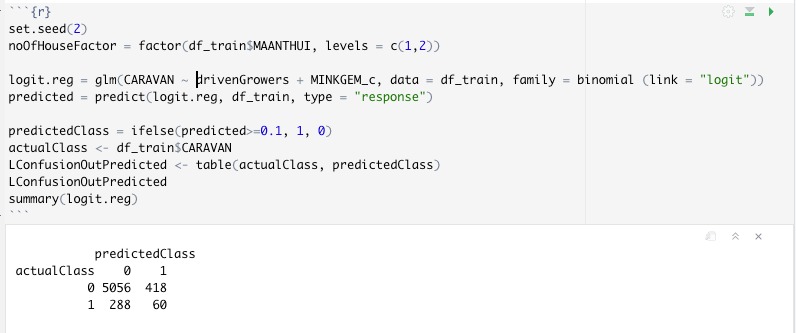
1. **Have you partitioned the data?**

Yes, we have partitioned our data with a 60% in testing and 40% in validation phase and then run some basic analysis on each set to understand what is happening with the data. The purpose of partition was to assess the accuracy of our model.

1. **Have you started running the model?**

Yes, we have started to run the model and are getting some initial results. Though the results are ok, but we are moving towards increasing the accuracy of the model by running each variable of interest by creating a confusion matrix.

Below is a glimpse of our confusion matrix:

****

1. **Remember you need to need to apply TWO techniques**

For now are trying to use the following:

* 1. GLM

**But in future we might also use multiple nominal regression**

1. **Any questions?**

So far so good

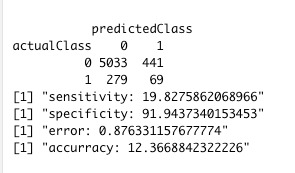
1. **Write a paragraph of the advances you have made for your project**

Our advances are as follows:

* We converted income as a categorical data to numeric data
* Identified characteristics of a potential customer who might by an insurance. Some of these are as follows:
  + Customers who live in 1 house buy more insurance than those who have more than 1 house
  + Age group also plays a role mainly between 40 – 50 years
  + Lower class and middle class families are buying more insurance
  + Family with grown-ups is also buying insurance
  + Households with size 3 or more are prone to buying more insurance
* We also did dummy conversions

After having don all this, we utilized the aforementioned to run a GLM model but found that a few of them were not significant such as age, household factor, avg family due to which response variable wasn’t changing much. However, we found that main customer type category, Driven Growers, was playing an important role and we could see the effect by removing it. Income was another variable which is having an impact on our response variable.

Below is a glimpse of our confusion matrix:



As it can be seen from the model that our sensitivity is low, we are still trying to figure out methods to increase it. Same applies to accuracy as well.

In summary, we are trying to put 3-4 models. First with our domain knowledge. 2nd we will use correlation matrix to figure minimum number of good predictors that are not correlated with each other and do trial and error method to figure them out.