# Problem:

Insurance companies rely on policy revenue for their profits. When customers cancel policies prior to a year, there is lost revenue. It is difficult to prevent this loss of revenue because we do not know who will cancel ahead of time.

# Uses:

An insurance company could potentially use this information for two purposes. First, they could use it to identify the leads that are likely to cancel and do not buy those leads. That money would be better spent of buying leads for customers who will be less likely to cancel.

Second, after a policy is purchased, identify those who are most likely to cancel. Then have targeted interactions with them aimed at preventing cancellations e.g. payment reminders, phone calls to ensure satisfaction.

# Data:

Data is available from a small insurance company. The data being used contains variables that are relevant to each lead in the data set: Demographic characteristics (gender, age, marital status, credit, geography), previous insurance information (insured vs. not insured, policy type, bi limit), car (make, annual mileage, number of vehicles), policy purchase information (same day vs. follow up sale)

# Approach:

The data analysis took place in three phases, which include data wrangling, exploratory analysis, and predictive modeling. In the data wrangling phase, the data from three different sources was combined. The lead data information included all of the information we know when a lead is purchased, the purchase data included all of the purchase information, and the cancel data gave us our target variable.

Another aspect of data wrangling was be examining missing data and determining the best course of action. Rows of the data set that contained multiple missing values were deleted from the data set. In other cases, the missing data indicated that there was no cancel or the customer didn’t want something revealed. In those cases, there was an other/missing category put into categorical variables. In cases of a numeric variable, the mean was using to fill the missing values.

# Limitations:

Insurance rates are always changing, so it is possible that anything we find could change in the future. This means that the model will have to be readjusted regularly to fit new data changes, additional insurance carriers, etc.

# Revised Approach:

My approach hasn’t changed with the data I have because I was generally familiar with the data coming into this project. I haven’t found anything that I have not expected. If my model doesn’t work after I’m finished, I might consider a model to predict revenue on a continuous scale.