# SYS 6018: Competition 4

# Xinyang Liu, Myron Chang

# xl9qw, mc7bk

**Reflection:**

**Who might care about this problem and why?**

Porto Seguro and other insurance companies may be interested in the predictions about whether a driver will file an insurance claim next year. This information is useful when insurance companies want to make pricing, client group targeting, and compensation plans for next year. For example, if a driver is highly likely to be in an accident next year, then the insurance company can advise him or her to pay more and get a more comprehensive plan.

For individuals, this problem may be of interest to those planning to buy an insurance plan, and wanting to have an idea about whether or not they should pay higher prices.

**Why might this problem be challenging?**

With a training set containing 59 columns and nearly 600,000 rows, along with a test set consisting of 58 columns and close to 900,000 rows, we did not have enough computing power to do calculations, build models, and make predictions using all the data. By subsetting the original data frame so that we can run our programs in an acceptable range of time, we have to sacrifice some degrees of accuracy in our models.

In addition, the values of the response variable in our training dataset are extremely imbalanced, with more than 570,000 observations being 0 and only 20,000 being 1, which makes it difficult to get a feasible and working model. Therefore, it is extremely difficult to make precise predictions.

**What other problems resemble this problem?**

The approach of this problem can be applied to any problem in which companies use various quantitative and categorical variables to create a model and predict the results of future marketing or pricing strategies. Predicting the probabilities of certain events, determining whether or not communities are highly likely to have criminal activities, and assessing advertising methods based on possible future sales are all potential problems that resemble Porto Seguro’s safe driver prediction problem.

With an unbiased multivariable dataset and a powerful model, we can learn a lot about predicting future probabilities. Companies could use the model and predictions to gain a better understanding about their business tactics and adjust accordingly, which could reduce their financial cost and improve their targeting strategies to acquire potential clients.