**SYS 6018: Systems Engineering Competition 3**

Team (Competition 4-8): Elizabeth Homan, Ali Zaidi, and Elaine Liu

**What is the problem?**

The Porto Seguro Kaggle Competition is focused on creating a model on anonymized predictors of car insurance information to find out if a person will file a claim in the next year. This problem is very feature-rich as the field of car insurance takes in information such as model of car, car age, driver age, experience, past accidents, and much more. In this problem, these predictors have been anonymized via transformations as the company is looking to use the modelling results in their own model to improve their business. The purpose of modelling this data is to figure out which drivers are most likely to get in an accident and submit a claim to charge these “bad” drivers an increased amount for insurance and charge “good” drivers with no claims less.

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

This challenge is posed by Porto Seguro, one of the larger auto insurance companies that operates in Brazil. The United States personal vehicle insurance premium market comprised of $186 billion in 2014, which is 35% of the entire property and casualty insurance premiums according to Trefis Research Group. The global market is orders of magnitude greater as auto insurance is found in most of the countries around the world. Therefore, being able to model this dense, feature-rich dataset and increase accurate predictions of what predictors contribute most to claim submissions is extremely important in a very large industry. The biggest issue for auto insurance companies is being able to determine which drivers are safest and which drivers are riskiest, to determine how much money to charge each customer as the price paid by each customer is variable dependent on many factors. To determine the price these customers pay, the company must predict which of these customers is going to submit a claim in the future and charge them accordingly. Currently, many drivers that are safer drivers get charged more and riskier drivers get charged less due to the model inaccuracy.

**Why might this problem be challenging?**

The car insurance market is very competitive as there are many companies vying to get customer business. There are hundreds of millions of drivers around the world driving hundreds of millions of different cars with thousands of potential predictors as claims can submitted for any number of reasons. Due to this, forecasting which drivers will file a claim in the coming year is very difficult and requires a lot of data on the vehicles, drivers, geographical location, etc, each which have hundreds of thousands of possible variables.

**What other problems resemble this problem?**

The overall insurance claim market resembles this problem – such as boat insurance, life insurance, home insurance, health insurance, etc. In each of these categories, there are thousands of predictors available and a similar binary classification exists on whether a customer will submit a claim or not.