As a competitive insurance agency, Porto Seguro wants to ensure that the drivers we insure do not carry more risk than we can afford. In order to better understand the potential of a client to file a claim, we were tasked with building a model that can make predictions, based on various identifying information and calculations, of whether or not a driver that we insure will file a claim. The data that we have available to build this model includes 595,212 observations that includes information on whether or not a driver filed a claim. For each observation, there are 57 independent variables (features) that include identifying information about the driver, the car, and various calculated fields.

Along with the 57 features that were present in the data, I created 8 additional features that I thought could potentially enhance the predictive power of the model. These features were created based on observed differences between the claim rates of the groups. The first new feature was a count of the number of missing values that were present across each observation. There were many missing values in the data, which would interrupt the modeling process if not taken care of. I chose to impute the column mean in place of the missing values. However, in order to not lose any potential insights, I created the new feature containing the number of missing values per observation. As an extension of this, I noticed a difference in claim rates between those with 4 or less and 6 or more missing values, so I created a binary feature that indicated to which of these groups the observation belong. Two additional features included counts of the binary features present for each observation. There was a difference in claim rates between those with 2 or less and those with 3 or more “ind” binary features, while there was no difference in claim rates for any counts of the “calc” binary features. The final features were based on how different these binary features are from one another, to determine if a larger difference has an effect on claim rates. For the “ind” binary features, there was a difference between those with 2 or less and those with 3 or more, while again no difference was observed for the “calc” binary features.

In order to select the features with the most predictive power, I ran a random forest model to determine variable importance. A random forest model ranks the variables based on how much permuting, or changing the order of and/or excluding, each feature decreases the accuracy of the model. If a feature is important, then permuting the feature should significantly decrease the accuracy of the model. Based on this, I then chose the top 60% of the features and used those to build my initial logistic regression model. I partitioned the data into training and test sets, in order to improve and test my model at different stages. After removing more of the insignificant/less significant features, I was left with a model that accurately predicted approximately 61% of the target cases. The area under the curve, which gives us a sense of the predictive power of the model, for my final model was approximately 63%, which means the model has somewhat decent predictive power.

As the final model is only 61% accurate, it does not completely satisfy the business problem of helping to eradicate the risk involved with insuring drivers. However, if the company is able to take on the remaining 39% of uncertainty, then our model can act as support for determining whether or not to insure drivers based on the various characteristics that were included in the model. Overall, the model is moderately good at making predictions about whether or not a driver insured by Porto Seguro will file a claim, however there is room for improvement.