Porto Seguro safe driver prediction

# Executive Summary

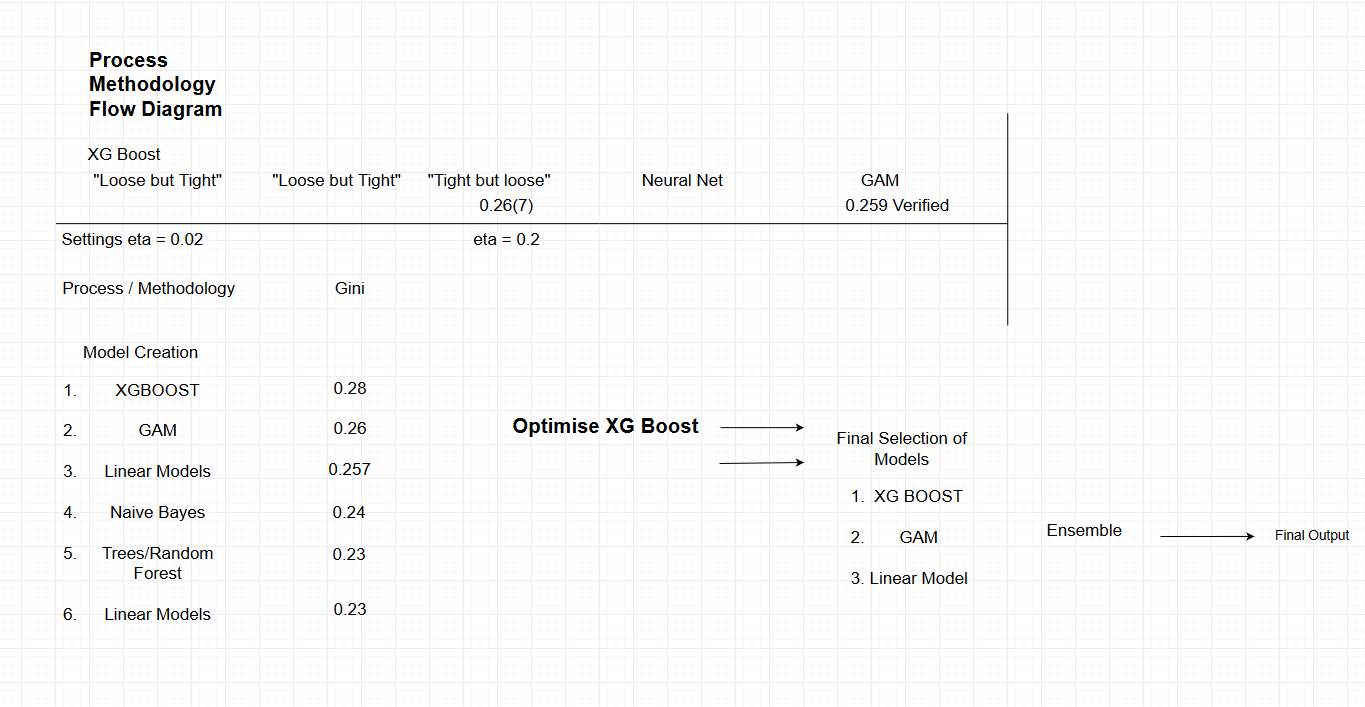
The average claim will set a car insurance company back £3000. Car accidents can be notoriously difficult to predict due to the inherent randomness of many accidents. Despite this we have been able to locate trends that can provide additional insight into individuals who may be involved in an accident. Based on the data you provided, we have harnessed several sophisticated and ground breaking modelling techniques including neural networks and gradient boosted models that have allowed us to identify high risk customers. Our unique approach involves creating an ensemble of several different model types in order to explain as much of the data as possible. We found that regional, individual and car information were useful in predicting claims however the calculated features within the data did not provide any significant prediction capabilities. Our model could potentially save Porto Seguro in excess of $200,000 a year going forward. With this new information Porto Seguro can alter prices according to the probability of claims and also refuse to take on customers that are likely to make them a loss.

# Exploratory Data Analysis

An initial look at the data confirms there are 1.4m records (in total between train and test files provided) with 59 features, the names of which were partially anonymized to comply with data privacy.As we alluded to earlier, car accidents are difficult to predict due to the inherent randomness of accidents and in general, how rare they are to happen to an average person. In our data set, the proportion of accidents is low – representing only 3.6% of the train data set. Furthermore, when dealing with human data features unknown information can be common as it is either not maintained in a consistent manor or the policy holder may not wish to provide, this dataset consists of 2.4% unknown values. Common approaches to handle unknown values are typically to take the average result, however we opted to create an additional feature storing whether an unknown value was present. This enabled us to test the hypothesis that when policy holders “withhold” certain information can improve prediction of the likelihood they will have an accident with the upcoming year.   
  
To quickly gain a general overview of features which yield predictive value, a simple Gradient Boost model was created. This model highlighted none of the 17 features containing the term “calc” added any predictive value to the model. These features were therefore removed from the list of candidate features to increase speed of model creation. Of note, our additional column which captures if any unknown values are present does appear in the top 20 features. Figure 1 below highlights the full top 20 features and highlights that two features: “ps\_car\_13” and “ps\_reg\_03” clearly stand out from the rest with predictive weights of ## and ## respectively.  
  
**{ insert Figure here }**

# Methodology

Due to the oversaturation of individuals who did not make a claim this makes developing a model that accurately predicts those that do make a claim much more challenging. As an example, one could create a model that predicts the correct result 97% of the time by simply saying no individual will ever make a claim. To overcome this, we chose to use an ensemble model technique which allowed us to exploit the benefits of many different models, each with their own unique take on the data set. In addition, ensemble models allowed us to apply different techniques to account for the oversaturation inherent within the provided data set. Techniques such as over and under sampling were applied, details of which can be found within the technical appendix.   
  
The list of models ultimately used in our final ensemble model are gradient boosting, generalized additive models and logistic regression (for a full list of all models tested and considered please see appendix X.   
  
Please see Figure 1 for a visual process flow of the methodology used to create and optimize our final model.

As we have used an ensemble model it can be less straightforward to get predictions from the model. For this reason, we have written a simple script that you can use which takes as input a data set of the same format that you provided us with and will then output a data frame containing the id’s of the customers and the probability of them making a claim. This will ensure you have no difficulties making use of the model despite its complexity. 

# Results

Of all the models considered gradient boosting yielded the best results. For this reason, we included three variants of this model type within the ensemble.

# Financial Benefit and Conclusion

The ability to identify high risk individuals will be a useful tool in increasing profit margins and attracting new customers. For every individual our model correctly identifies this will save roughly $2000 by not taking on high risk individuals or increasing premiums for those identified. This figure is calculated using an average of $500 for car insurance per annum in Brazil and an average claim of $2500. Figure 2 displays the potential annual net savings for Porto Seguro by identifying high risk customers at different probabilities of having an accident. For an outset of just $25000 you will make a return on investment of 800% in the first year alone. We have not only conducted analysis to build a predictive model but also looked at the insurance sector in Brazil to see how our model can best work for you. The insurance market is still growing in Brazil with only a 3.5% penetration as of 2012. However, the market is growing at a rapid pace. Through using our model, you will be able to identify low risk individuals and thus offer them competitive rates ensuring that more new customers will sign up with you rather than going to the competition. This will further increase the financial value of our model beyond that of saving on claims.