# PORTO SEGURO SAFE DRIVER PREDICTION

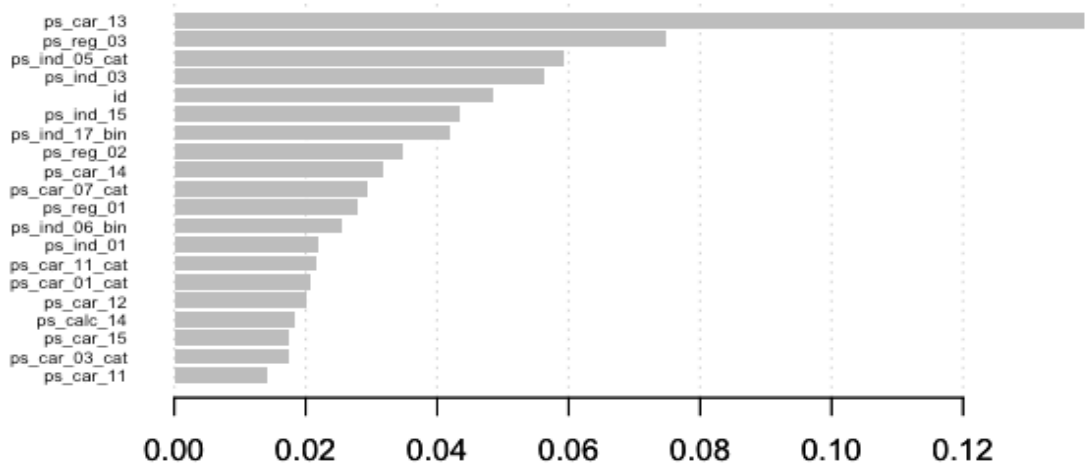
## EXECUTIVE SUMMARY

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 $260,000 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 manner 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 the likelihood they will have an accident within the upcoming year will increase, which aided prediction.

*Table 1: The top 20 features within a simple gradient boost model, the x axis displays the weighting of the feature in the model*



To quickly gain a general overview of features which yield predictive value, a simple Gradient Boost model was created. This model highlighted that only 1 of the 17 features containing the term "calc" appeared in the top 20 features. These features were therefore removed from the list of candidate features to increase the speed of model creation. Figure 1 above highlights the full top 15 features and highlights that two features: "ps_car_13" and "ps_reg_03" clearly stand out from the rest with predictive weights of 0.12 and 0.075 respectively.

# METHODOLOGY

Due to the oversaturation of individuals who did not make a claim, developing a model that accurately predicts those that do make a claim was very challenging. As an example, one could create a model that is 97% accurate by simply saying no individual will ever make a claim. To overcome this, we chose to use a ground breaking 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 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. We meticulously tuned each model to ensure that we achieve the best possible accuracy.

Figure 1 demonstrates 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.

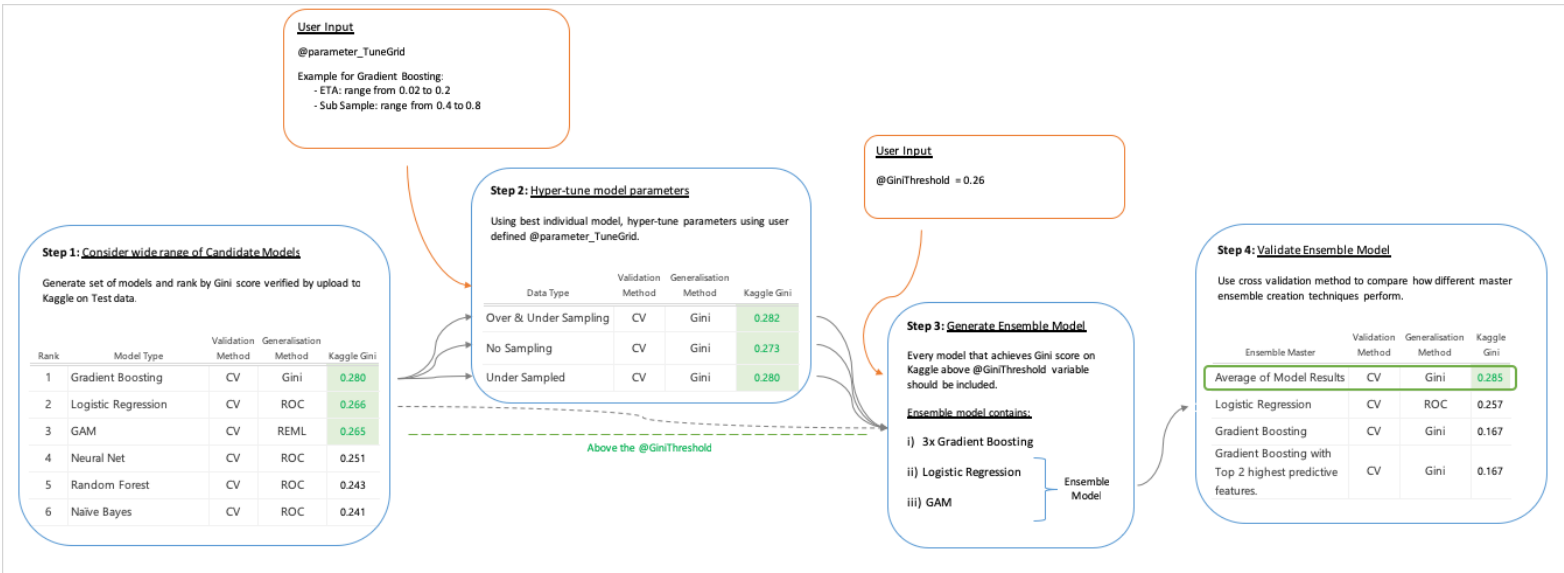



*Figure 1: Process flow of ensemble model design*

## Results and Financial Benefit

Of the large selection of models considered we found gradient boosting yielded the best prediction results. We included 3 variants of this within the ensemble. On top of this we added two other models which captured different aspects of the data, a generalized additive model and a logistic regression. These produced a broader indication of the underlying trends within the data. As a result, our model produces dependable results with no noticeable drop in accuracy when applied to new data.

The ability to identify high risk individuals will be a useful tool in increasing profit margins and attracting new customers. Porto Seguro can save roughly $2000 for every high-risk individual that our model identifies. This could be through refusing to take them on or by offering them higher premiums. This figure is calculated using an average of $500 for car insurance per annum in Brazil and an average claim of $2500. Figures 2 and 3 display the potential annual net savings for Porto Seguro by identifying high risk customers at different probabilities of having an accident. For a relatively small license fee of $25,000 Porto Seguro could save in excess of $260,000 per year. This provides an 800% return on investment within the first year alone. This is based on a sample of 600,000 customers however could be scaled up to the total 1.4 million customers to return an estimated saving of $600,000 per year. 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 (*International Monetary Fund, 2012)*. However, the market is growing at a rapid pace. Through using our model, you will also 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.

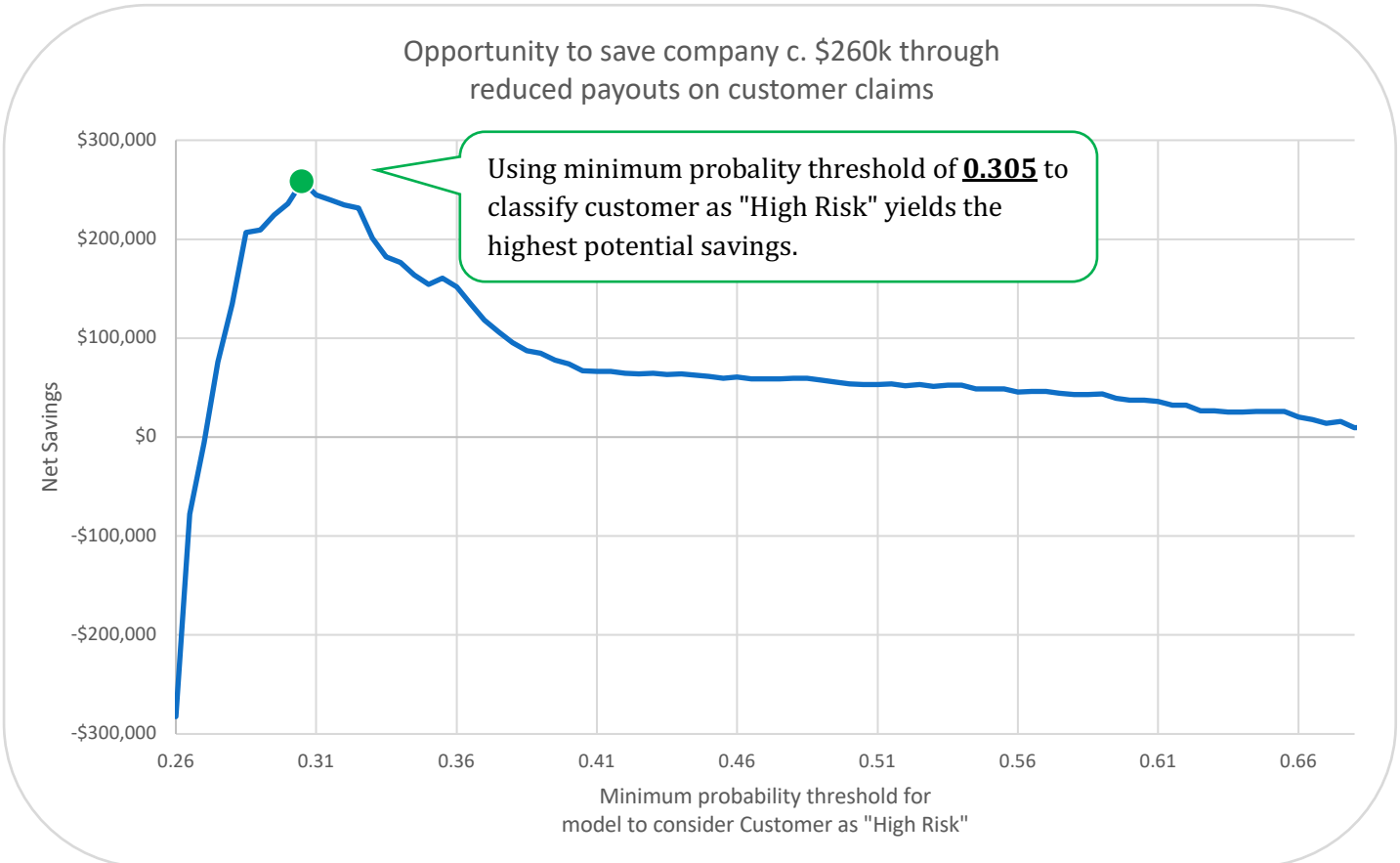



Figure 2: Total savings at different thresholds identifying high risk customers - savings based on a sample of 600,000 customers

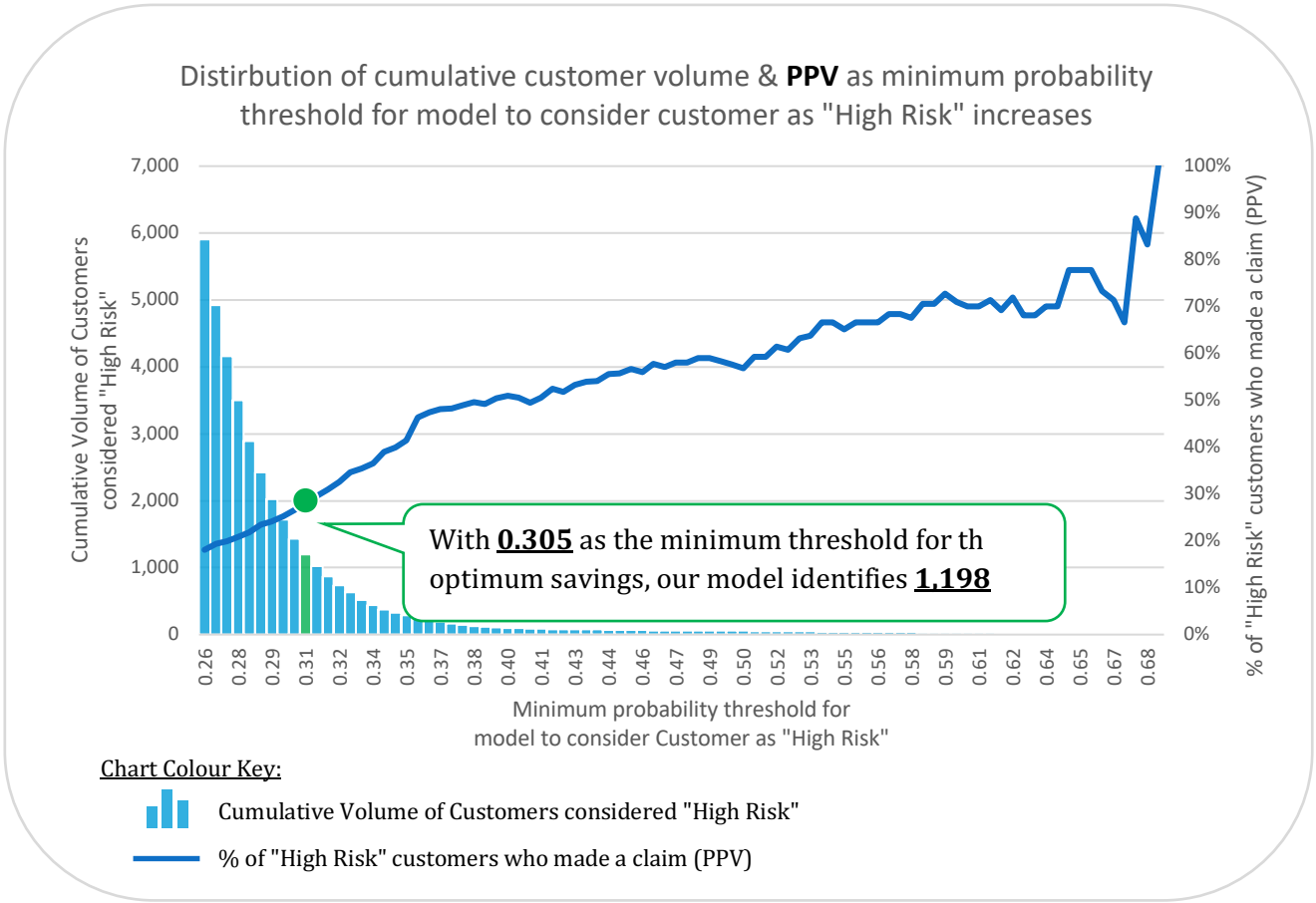



Figure 3: Number of customers identified as high-risk vs those that actually made a claim

## Conclusion

Our modelling can provide numerous advantages for Porto Seguro that will give you an edge in a competitive and growing market. This can be through adjusting premiums, rejecting risky applications and attracting new low risk customers with competitive rates. Using cutting edge methods, we have built a model that is capable of capturing trends in a turbulent and often unpredictable market. For a relatively small $25,000 investment Porto Seguro could benefit for years to come.

## References

International Monetary Fund. Monetary and Capital Markets Department, (2012). Brazil : Detailed Assessment of Observance of Insurance Core Principles of the International Association of Insurance Supervisors. USA: INTERNATIONAL MONETARY FUND. doi: https://doi.org/10.5089/9781475591354.002

Kaggle, (2019). Porto Seguro safe driver prediction data. Available from: https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data