

Surya Prakash Bhoi

Data Driven | Data Science | Data Quality


LinkedIn: <https://www.linkedin.com/in/suryapbhoi/>

GitHub: <https://github.com/suryapbhoi>




Driver Insurance Claim Prediction for Porto Seguro S.A.


The Problem

- Porto Seguro, one of Brazil's largest auto and homeowner insurance companies is currently facing the problem of inaccuracies in car insurance company's claim predictions, which results in increased cost of insurance for good drivers and reduces the price for bad ones
 - The “insurance premium amount” – Ideally, safe drivers should be charged a lower insurance premium for being more responsible drivers. And unsafe drivers should be charged higher insurance premium amount for being careless or insensitive towards the traffic rules, and causing accidents on the road
- 


The Solution

- The objective was to leverage the driver's historical data and come up with a solution which finds out whether a driver is likely to file a insurance claim next year
 - The solution can be translated as:
 - Low risk customers (i.e., drivers who are unlikely to file a claim next year) are charged lower premium amount
 - High risk customers (i.e., drivers who are likely to file a claim next year) are charged higher premium amount
- 

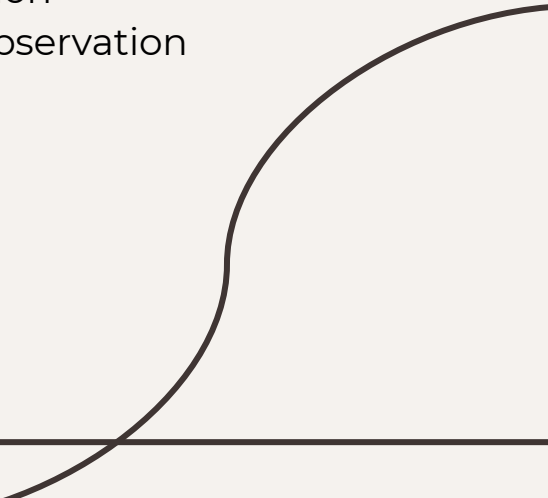
The Challenges

- The feature names in the dataset are “masked” instead of real feature names. Hence, our analysis of the dataset will not be more interpretable. For e.g., we could not conclude observable analysis results like, who files more insurance claims
 - Commercial vs Personal vehicle drivers
 - Male vs Female drivers
 - Young vs Elder drivers
 - The dataset size was big
 - Train dataset was 115 MB, having 595212 observations
 - Test dataset was 172 MB, having 892816 observations
- 

The Approach

- This being a Machine Learning classification task, the below algorithms were used to determine the target:
 - Logistic Regression
 - Naïve Bayes
 - Decision Trees
 - Random Forest
 - Gradient Boosting Tree Classifier
 - Stacking Ensemble Classifier
 - Multi-Layer Perceptron Classifier
 - AUROC and Log-Loss were the metrics used to compare the various models, and determine the best one
- 

The Predictive System Delivered

- The ElasticNet Logistic Regression was the best among all the models we tried for the solution
 - AUROC of 0.617
 - Log Loss of 0.153
 - The model was deployed as a web application (using streamlit cloud platform), to be used by associates and managers in any browser
 - Predictive system which predicts for a single observation
 - Predictive system which predicts for more than one observation stored in a csv file
- 

Thanks

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**