Project Summary - Porto Seguro’ s Safe Driver Prediction

Team22 - Zihan Wan ＆ Rongjing Huang

---

1. Project Introduction

Porto Seguro Insurance wants to use machine learning models for risk identification to more accurately predict the probability that an insured driver will initiate an auto insurance claim in the following year, thereby adjusting its auto insurance pricing strategy to reduce payout risk and expand its customer market.

1.1 Project Link

This is a competition in Kaggle:

https://www.kaggle.com/competitions/porto-seguro-safe-driver-prediction

1.2 Project Background

- Porto Seguro Insurance Company's product line (for individuals/businesses/property/vehicles)

- Introduction to auto insurance from Porto Seguro Insurance Company (basic content/personalization)

2. Project Goals

2.1 Problem Statement

- For insurance companies
  - Low accuracy in identifying the risk of customer claims, lack of clarity in the user profile of customers, resulting in large losses in insurance companies' claims, low customer stickiness and difficulty in user acquisition.
- For customers
  - Solidified or unreasonable pricing of services, resulting in low willingness of customers to enroll or renew their insurance.

2.2 Importance

- For insurance companies
  - Based on the model to get the key characteristics of the customer, saving the cost of customer research and facilitating the rapid assessment of the risk of new user claims.
  - Assessing claim risks based on customer information to provide accurate and reasonable pricing plans can improve customer experience while reducing the company's claim losses, which helps reduce customer churn and develop the customer market.
- For customers
  - Customers can enjoy customized solutions with reasonable pricing, reduce unnecessary expenses and get more cost-effective service content.

3. Data Description

- **train.csv**
  - Data Size：
    - 595212 rows × 59 columns
  - Data Type：
    - int64 - ×49（including id、target）
    - float64 - ×10
  - Serial No.：`id` - User id
- **test.csv**
  - Data Size：
    - 892816 rows × 58 columns
  - Data Type：
    - int64 - ×48（including id、target）
    - float64 - ×10
  - Serial No.：`id` - User id

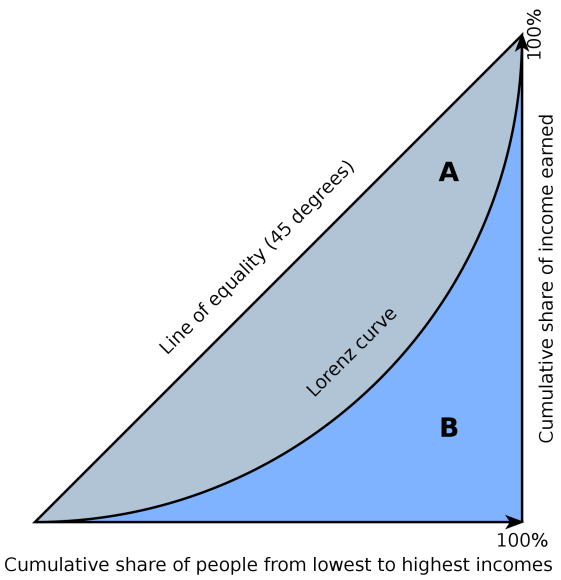
- Target：`target` - 1(Claim)，0(No Claim)
- Features（×57）：
  - `ps_ind_xx` - Individual Info（×18）
  - `ps_reg_xx` - Region Info（×3）
  - `ps_car_xx` - Car Info（×16）
  - `ps_calc_xx` - Calculated Info（×20）
- Feature Type：
  - Binary（Suffix with bin） - ×17
  - Categorical（Suffix with cat） - ×14个
  - Continuous / Ordinal - ×26
- Note：
  - Missing values are indicated by -1

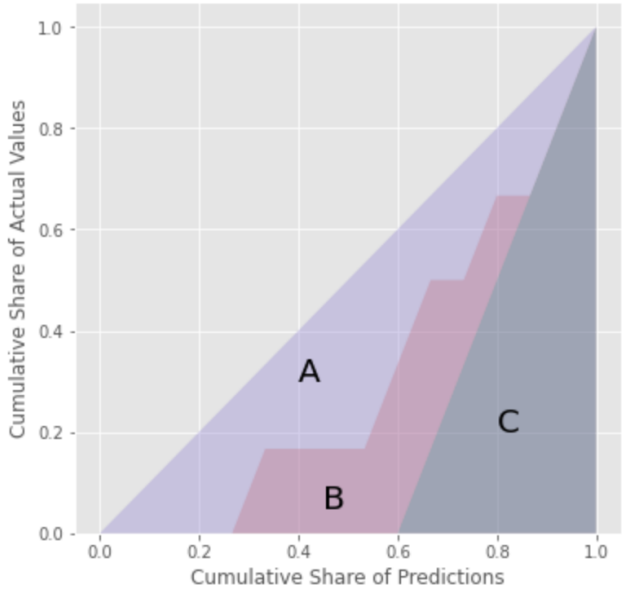
- No target
- Features（×57）：
  - `ps_ind_xx` - Individual Info（×18）
  - `ps_reg_xx` - Region Info（×3）
  - `ps_car_xx` - Car Info（×16）
  - `ps_calc_xx` - Calculated Info（×20）
- Feature Type：
  - Binary（Suffix with bin） - ×17
  - Categorical（Suffix with cat） - ×14
  - Continuous / Ordinal - ×26
- Note：
  - Missing values are indicated by -1

4. Evaluation Methodology

4.1 Scoring Metric

- For unbalanced data, where the prediction of a very large magnitude of data in a dichotomous problem leads to dilution of the prediction of a very small magnitude of data, we cannot use traditional accuracy to measure the goodness of the model.
- The confusion matrix can measure the prediction accuracy and coverage of the two types of data through precision and recall and f-score respectively, but it cannot be synthesized into one indicator to reflect whether the prediction of the two types of data together is good or bad.
- So, the prediction results of the model are evaluated by the Normalized Gini Coefficient.
- In this project, the Gini coefficient indicates the effectiveness of the model in distinguishing between "bad" drivers, who will claim in the future, and "good" drivers, who will not claim in the future.
- The Gini coefficient ranges from 0 to 0.5, and the normalized Gini coefficient is adjusted with the theoretical maximum so that the maximum value is 1.





4.2 Principle Implementation

- Sorting the values of prediction from smallest to largest and adjusting the order of actuals accordingly.
- Accumulate and plot cumulative line graphs (i.e., Lorenz curves) for the adjusted sorted actuals.
- Normalize the x and y axes (i.e., convert between 0 and 1) and plot the 45° diagonal;
- Calculate the area of A to obtain the Gini coefficient a.
- Theoretically, the maximum Gini coefficient can be obtained when predictions and actuals are perfectly matched.
- Therefore, the values of actual are sorted from smallest to largest, summed and plotted, and the area of A+B is calculated to obtain the maximum Gini coefficient b.
- Finally, a/b is calculated, which is the normalized Gini coefficient.

5. Project Process

- Read Data
- Build metadata
- Outlier detection
- Missing Value handling(Multivariate imputation/Univariate imputation)
- EDA(Discrete feature analysis/Continuous feature analysis)
- Feature Engineering(Encoding/Feature Transformation/Feature Enhancement/Feature Selection)

- Data Modeling(Decision Tree/Logistic Regression/XGBoost)

6. Project Results

Finally, our model obtained the following score, which is good given the highest score of 0.296 on kaggle.

```
Normalized gini coefficent for the entire training set ( combined ):
 final normalized gini score = 0.2822374207618298
Wall time: 7min 29s
```