1. The project's domain background

   This project is based on Porto Seguro's Safe Driver Prediction Competition on Kaggle[1].

   Porto Seguro is one of the biggest auto insurance company in Brazil.

   Porto Seguro has been working with machine learning over the last 20 years, but now they rely on Kaggle community to build a new fair algorithm.

   A common problem for an insurance company is to predict if a person whether a person will claim insurance in the coming years.

   Inaccurate evaluations cost a lot of money, so being able to predict if a new insurance is a good investment or not, it's important.

2. A problem statements

   It doesn't seem fair to pay more for an auto insurance, since you have been driving safe on streets.

   So, the project's goal is to identify good drivers through inputs, so they can pay less for auto insurance, and penalize the bad drivers.

3. The datasets and inputs

   There are 2 files for this competition:

   - Test File[2]
   - Train File[3]

   The test file is to evaluate your algorithm and the train file to is build your algorithm model.

   Test file has 595.213 records with 58 inputs.

   Most of inputs are anonymized since information are related to customers.

   - First column: customer ID,
   - Second column: target (1 - claiming insurance, 0 – no claiming),
   - Third column to twentieth column (18 inputs): personal information,
   - Twenty-first column to twenty-third column (3 inputs): region information,
   - Twenty-fourth column to thirtieth-octave column (15 inputs): auto information,

[1] https://www.kaggle.com/c/porto-seguro-safe-driver-prediction#description

[2] https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/download/test.7z

[3] https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/download/train.7z

- Thirtieth-ninth column to fiftieth-octave column (20 inputs): calculated feature.

4. A solution statement

   A possible solution is to train the train inputs to create a supervised learning model. It will be used Decision Trees and SVM algorithm.

   It’s known that the Competition should be based on probabilistic classification, for example Naïve Bayes or Gradient Descent.

   The supervised learning model will be tested through test file, and then will be checked the results (Number of targets that match).

5. A benchmark model

   Since it has been provided the test file, the benchmark model will be the test file.

6. A set of evaluation metrics

   The evaluation metric is the target (Percentage of correct target predictions).

7. An outline of the project design
   a. Create File in IPython
   b. Import Data and library
   c. Format the data so it fits the library
   d. Train the algorithm
   e. Evaluate the results
   f. Upload Results