**Project 1:** Predict Toxicity of online comments

This is a project hosted on Kaggle at the following URL

<https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data>

**WARNING:** The sample data are extracted from real world and contain words, sentences and phrases that may be offensive to some. If you are sensitive and likely to get offended, please **DO NOT** work on this project. Data here are provided entirely for learning purposes and obtained directly from the Kaggle website, author **does not** take any responsibility on the content.

You are provided with a large number of Wikipedia comments which have been labeled by human raters for toxic behavior. The types of toxicity are:

* toxic
* severe\_toxic
* obscene
* threat
* insult
* identity\_hate

You must create a model which predicts a probability of each type of toxicity for each comment.

**Data File descriptions**

train.csv - the training set, contains comments with their binary labels

test.csv - the test set, you must predict the toxicity probabilities for these comments. To deter hand labeling, the test set contains some comments which are not included in scoring.

sample\_submission.csv - a sample submission file in the correct format

test\_labels.csv - labels for the test data; value of -1 indicates it was not used for scoring; (Note: file added after competition , in this class you may use it as a reference)

Solved using LogisticRegression

<https://www.kaggle.com/tunguz/logistic-regression-with-words-and-char-n-grams>

**Project 2:** Predict whether a Loan applicant will miss payments (default)

<https://www.kaggle.com/c/home-credit-default-risk/data>

Review the web page and understand various files. *application\_train.csv* has pre-classified data for training. Column label TARGET has the status whether and applicant is likely to miss payments.

Description of all the fields are explained in HomeCredit\_columns.csv file

Review the data correlation diagram provided in the web-page and understand how each file is related. Note: These files are very large, sometimes it would not be possible to open with excel, in such event please use text editor such as Notepad++

Project:

1. Implement an AI model to predict whether an applicant is likely to miss payments. And derive the accuracy of the model.
2. Use application\_test.csv and predict what applications are likely to miss payment and list the application ID of those who will miss payments.
3. Include past home loan installment payment history and credit card payment history as new features. Re-train the model and observe the accuracy
4. Introduce PCA and experiment to see how it affect accuracy

Hint for number 3:

Consider file installments\_payments.csv. This file represents client’s installment payment history for past loans.

As an example: consider following fields

|  |  |
| --- | --- |
| DAYS\_INSTALMENT | Installment due day |
| DAYS\_ENTRY\_PAYMENT | Actual day it was paid |
| AMT\_INSTALMENT | Due Installment amount |
| AMT\_PAYMENT | Amount actually paid |

From this you may derive payment behavior (pattern) of this client.

Pay\_day\_delta = DAYS\_INSTALMENT - DAYS\_ENTRY\_PAYMENT

Number of Late payments = SUM (pay\_delta + grace\_period < 0)

Pay\_amount\_delta = AMT\_PAYMENT - AMT\_INSTALMENT

Number of deficit payment = SUM( Pay\_amount\_detla < 0)

This is only one possible way, you may use other representations.

Following URL gives some additional explanations that may be useful

<https://www.kaggle.com/pavanraj159/loan-repayers-v-s-loan-defaulters-home-credit>

**Project 3:** Predicting amount of future financial needs of customers

This is Regression problem. Problem is copied from Kaggle. In summary, you are required to build a model that predict size of potential future financial needs of customers. Having ability to foresee customer needs are perceived as having a friend in the business.

More details of the problem is explained at

<https://www.kaggle.com/c/santander-value-prediction-challenge>

There are two files

train.csv, that contains classified data

test.csv contains customer data for prediction

Colum names have ID, Target and series of 9 digit hex numbers. ID is the ID of customers. Target is the future financial needs. Column with Hex numbers are the features. They represent some sort of financial behavior and anonymized for this purpose.

Design a Regression model to predict customers future financial need

Sample solution is provided here

In this solution columns with single value or empty are dropped. This is an optimization and you may try it without such optimization and observe the accuracy (Mean Square Error)

<https://www.kaggle.com/amarjeet007/random-forest-feature-engineering-lightgbm/notebook>