**Cornell University**

**ECE 4200 Introduction to Machine Learning**

**Model report: Kaggle Competition**

**December 2020**

**Nicolás Oriol Guerra (no227)**

**1. Introduction**

The goal of this competition is to achieve the best possible performance in a model to predict different font types. As features, we are provided with inherent characteristics of observations of each of the different fonts. Overall performance will be measured by accuracy on provided testing data.

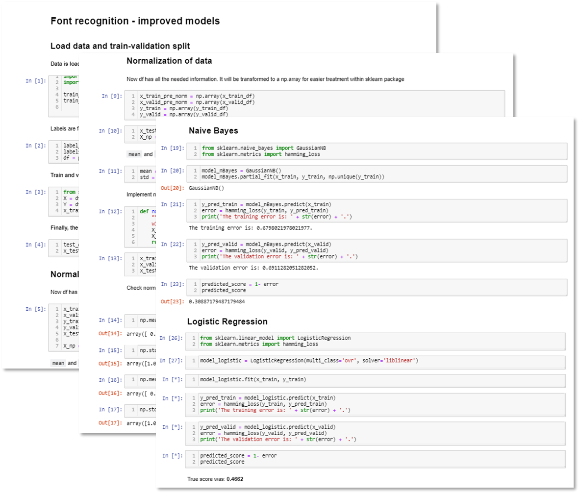
**2. Methodology**

The problem will be approached in two different ways: fitting of base models enforced by competition guidelines and fitting of improved models.

Base models are naïve bayes and logistic regression. Code and performance will be recorded in the `BaseModels` notebook. Improved models are different types and combinations and parameter settings of neural networks. Their performance is recorded in ÌmprovedModels` notebook.

Before starting with any model building, the data is normalized and split into training (70%) and validation (30%) sets. The validation set is used to predict model performance in the test set. Its predictions have been found to be very accurate and overall performance can be estimated with just the validation set.

**3. Models, performance and prediction scores**

The general approach to the problem consisted of three main steps: computing base models, training neural networks and performing PCA on the dataset before training the models.

Regarding the performance of base models, it can be said that it is suboptimal. Naïve bayes had a predicted performance of 0.309 and the true score was 0.323. For the logistic model, these numbers were 0.470 and 0.466 respectively.

In the case of neural networks, several approaches were taken. First, parameters were tries at random which initially brought significant performance improvements (up to 0.669 predicted score).

A bottleneck was experienced when trying to train several different neural networks using a for loop and varying parameters. The reason for this was computing power. For this reason, only certain parameter settings were iterated over, thus creating 40 networks to compare and raising the need to use random search cv. Afterwards, deep neural networks were tried by hand by using Google Collab notebooks as additional computing power. Performance for each of these 14 tests is found in `NNetworkTests` notebook.

Finally, PCA was

**4. Conclusion**

This practical case is an example of the anchoring bias explained in the heuristics section of