Individual Assignment Description

January 29, 2019

1 Dataset

The experimental dataset we are going to use is the HR Analytics Dataset. It includes explanatory variables of around 15k employees of a large company. The goal of the case study is to model the **probability of attrition** (employees leaving, either on their own or because they got fired) of each employee, as well as to understand which variables are the most important ones and need to be addressed right away.

The results obtained will be helpful for the management in order to understand what changes they should make to their workplace to get most of their employees to stay.

For more details on the dataset and the task check in Kaggle

2 Goals

The individual assignment goal is to work on a complex classification problem with machine learning tools covered (up to the deadline of this assignment) in ML-I and ML-II. The work will include the following sections:

1. **Exploratory Data Analysis**. The goal is to provide some insights on predictors and data structure that will guide the feature engineering work. As a result of this stage the dataset will be transformed by removing errors, imputing missing values and leaving only those features that are really informative.
2. **Baseline**. At this point it is important to run a basic model over the raw dataset (resulting from previous stage) and check what is the model performance on the classification task we want to achieve.
3. **Feature Engineering**. This stage will select and construct the variables that will be used in our modeling task. The former will analyze the effect of removing predictors upon some of the criteria covered in class. The later constructs new features and checks the effect on a baseline model.

Steps 1 to 3 are **iterative**. You must show with data whether your decisions are working towards a better model or not, to finally make a recommendation on what are the best features and the best model. To do so, please use feature construction, extraction and selection methods covered.

The models recommended are: linear and logistic regression and lasso or ridge regression (in case you use these, you must tune the parameter λ to find the best possible value).

2.1 Validation metric

The metric proposed to validate your work is the **accuracy**, that will be in-depth covered by the session #3 in class.

3 Deadline

The proposed deadline for this assignment is: **February 10th @ 23:59:59**.

4 Deliverable format

The deliverable will be uploaded to the IE Campus section for the assignment. It will consists of **two separate files** (mandatory):

1. A PDF file (named with your full name) with the description of your work. You must summarize in a **maximum of 5 pages** all the decisions taken along the machine learning process that ended up in a model validated with a given score. You can add any plots that will help to understand your decisions.
2. A ZIP file (also named with your full name) with all the files you used to work on the problem. It must contain a Python notebook, an R markdown or a DSS project file.

Submission example:

```
jesus_renero.pdf
jesus_renero.zip
```

5 Grading

This work will be graded with a score between 0 and 3, according to two simple criteria:

1. ML process comprehension (80%): This is the most important part, as it will score the quality of the work from what you're able to explain through the PDF summary, and how have you decided to build and run iterative ML pipeline to come up with different scenarios and models, to finally decide on what is the best set of features to start the modeling part.
2. Model performance (20%): A tiny fraction of the grade will be based on what you've been able to achieve with the problem. This score will depend on your rank among the results from all the class.

6 Mandatory submission parts

You all will work on the same dataset and with the same goal: define what is the best feature engineering strategy to start the modeling part. Your results will be based on the assumptions you make and the number of different techniques that you experiment with.

Following you can find a list of sections that MUST be present in your design:

6.1 Data loading:

Read functions from either an URL or PATH in your computer. The process must recover the entire raw dataset.

6.2 Data preparation

You must clearly separate the data preparations part from the raw data, you can call this method as many times as needed in the future to reset your changes. You must include:

- Hunt and impute NA's, if applicable.
- Refactor to the proper data type
- Scale and Skewnees must be fixed.
- Build a method to split into training and test (hold out) datasets.

6.3 Baseline

Baseline your model at this point. Once you've your features clean, baseline a model (you decide which one to use for this classification problem).

6.4 Feature engineering

Start with feature engineering process

- Decide on what will be your cross validation strategy and decide whether you implement your own or use any of the existing libraries in R, Python or Dataiku to do so. Apply this strategy robustly across the entire process of feature engineering to check your partial results after each step.
- Apply some Feature Construction idea to your dataset.
- Deal with outliers. It's part of this process because you must assess if removing them make your model stronger or weaker as part of the iterative process. Sometimes outliers are simply outplaced entries.
- Apply Feature Selection: Check if you can apply filtering methods (different to numerical and categorical features) to reduce the number of features. Apply also a wrapper method to check if your model improves.
- **Optionally** apply regularization in lasso or ridge regression, and try to find the optimal λ.
- **Optionally** apply Feature Extraction methods from Deep Feature Synthesis or Genetic Programming libraries.

6.5 Final metric

Report what's the score of your model with the final set of features.