# ­­Rubric Data Engineer and Data Scientist individual assignment I

|  |  |  |
| --- | --- | --- |
| **Assessment Criteria –** Data Engineer and Data Scientist individual assignment I | | |
| Studentnumber: | Studentname: | Grading: |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Insufficient 0 - 25 points | Marginal 26 - 55 points | Good 55 - 75 points | Excellent 75 - 100 points |
| *Data discovery* | The student uses only the provided dataset and has little understanding of its content | The student has added a minimum of 10 hand written reviews and has turned the dataset into a data frame  And the student has scraped Tripadvisor using the sample script.. | Additional to *Marginal*:  The student has scraped and labeled more than the minimum of 10 reviews from more than one hotel booking site and has turned the dataset into a data frame. | Additional to *Good*:  The student has scraped and labeled more than the minimum of 100 reviews from several hotel booking site and has turned the dataset into a data frame. |
| *Data preparation* | The student can barely turn the provided dataset into a usable dataset of labeled data. There is no live connection with a SQL database. | The student can turn the dataset into a usable dataset of labeled data and perform some additional cleaning if needed. There is a live connection with a SQL database, no parametrized queries | Additional to *Marginal*:  Moreover parametrized querying is part of the script | Additional to *Good*:  No embedded SQL is used in the script only stored procedures are used. More over some advanced cleaning had to be done |
| *Model planning* | The student has no idea about different models to be used for datascience | Student only knows to describe the models involved in the script. But has no ideas about the pro’s and the cons of the 3 models | Additional to *Marginal:*  The student can explain the ranked accuracy of the 3 different models. In short, why is a model better than another? | Additional to *Good*:  Student has done some research on classifiers, and can use arguments for using a particular one beyond the mandatory literature |
| *Model building* | Student cannot explain any of different statements in the code used to build a classifier. The dataset is not splitted into a training and a test set | Student can explain only the basic statements in the code behind only one classifier. The dataset is splitted into a training and a test set | Additional to *Marginal*:  Student knows how to explain all the ins and outs of the pieces of code involved. In particular how to succeed in improving the overall accuracy | Additional to *Good*:  Advanced tweaking of the parameters involved in the used classifiers has been used |

Minimum requirement for a pass ( i.e. grading ≥ 5.5):

* At least 3 out of 4 are *Good*
* None of them is *Insufficient*