# 28.02.2020, 10am15 to 11am15, 1-to-1 with Michael Notter

* **Chapter 4, Feature Discussion** : Are there some more plots that are missing or would make sense?
  + Introduce additional attribute “Missing” per feature. This should reduce bias of -1.0 and -0.5. Alternatively, set to -0.001. The size of the negative value could explicitly influence neural networks as they use thresholds for firing of a neuron. Or scaling?
  + Show correlation matrix heatmap.
  + Play with tsne of sklearn. Use a 2-dim tsne model.
* **Model.csv** : The simulation data is stored in .csv files. Have look at them for discussion.
* **Decision Tree Model** : I search for the best DecisionTree model to be at the maximum value of the validation accuracy before it decreases again.
  + With the balanced data set, there is no maximum, though.
  + There is no regularization for DecisionTrees, neither.
  + => ?
  + Is cross-validation needed?
    - You can play around with random\_seed, instead. You can also mention that cv would make the statistics robust.
    - It would be better to have a train, validation and test set, though!
    - Synthetic data!
  + The curves look like oversampling could help.
  + Instead of regularization, modify : leave size, only split when 10 elements are split, …
  + In the graphics, you can enter the value of the DummyClassifier as a line. This is called Chance level. But the validation curve is well above that line. All is good.
* **RandomForestClassifier** : How can I present the results and discuss them?
  + I am looking for a graphical analogy to DecisionTree.
  + Try with feature\_importances\_.
* **SVC** : How can I present the results and discuss them? (See RandomForest)
  + Find out, which books are always wrongly classified with all models?
* **Neural Networks** : Discuss results of NN, depending on grid search parameters.
  + Is l2\_alpha = 0 allowed?
  + My validation score does not exceed 99.9. What can I do?
  + I can raise the validation score slightly (to 99.89) with an additional second hidden layer. Does this result justify the extra effort?
  + I have to extend my epochs very much to reach a flat accuracy curve. Why?
  + The validation accuracy curve is higher than the training curve. Does that make sense?
  + Scale all plots the same way in the graphs.
  + Play with the architecture. That is the only promising opportunity.
* **Anything missing**?

# 14.01.2020, 09am to 10am, 1-to-1 with Michael Notter

* **Feature Matrix Generation** : Is it OK to work specifically with Python library *textdistance*.?
  + => To use textdistance is OK. Nothing new is needed.

To decide on the quality of an implementation, look at the homepage info where the library was found: The number of stars given to a library, how many people have voted for a library. And how many people have been implementing for the library and when has library had last commits? Look at the Releases. => This will give a feeling on its quality.

Presence in pip is not a quality label.

* + What is recommended on validating the results of the library?
  + => Compare the results of 1 to 3 metrics. Take an simple sample data set, like ‘Hello world’ and show the results with this data set. => Into Appendix.
* The Dummy Baseline is missing with the model comparison!
* Nuances of the models with an accuracy of 99,x% can be increased by using log(1-p).
* **Models** : Is it recommended to use oversampling of the undersampled features?
  + => Undecided. Test and find out.
  + => DecisionTreeClassifier and SVC know class\_weight = ‘balanced’. Use this, instead.
  + => Neural Network with keras knows class\_weights (ratio or percent or absolute?) as well in model.fit(). Try around with this.
  + Maybe sampling will still be needed and might make sense.
* IN[2], Neural Network Model => Careful with absolute values. If number of records change in goldstandard, the sampling might fail. => work with dynamically calculated len() amounts.
* **Neural Network** : What is the recommended network architecture?
  + => Having more layers, the network learns more interactions.

It might make sense, to expand the features with different metrics on same attribute. With this, the network might learn more nuances.

* + What are reasonable parameters to start with?

=> Topology makes sense. Try increasing topologies, though : 8 -> 50 -> 100 -> 2.  
=> Droput : try with and without. Remember : with Dropout of 50%, maybe person and title will be dropped and the remaining features for comparison are poor, then.

=> Use softmax, not sigmoid in output layer (2 classes)  
Use AdamOptimizer, not SGD.

* + How to do grid search with keras library?
  + => Add a chapter on Feature Matrix EDA : Discuss results of feature matrix. Use .describe() for a first analysis. Discuss all 12 features e.g. in plots with colour codes (duplicate, non-duplicate). (This might influence decision, if you want to add new redundant features.)
  + => Depending on the feature distribution, Scaling might make sense. Use separate scalings for different features.
* Does **structure of the project** match and is it complete?

=> Looks generally fine. The discussion of the results is important, though.  
Look at some sample pairs that are outside the diagonal of the confusion matrix, compare the models on these outsiders and discuss advantage and disadvantage of models. => ScikitLearn knows VotingClassifier. -> Get feeling for models.

* Am I allowed to still change and complete the proposal?
  + <= The calculation of the test data amount is wrong.
  + => Yes.
* I do not have specific marking of Observation, Decision, but I write more a continuous text, like in a book. Is this style fine?  
  => Fine.

Ein Bild, das Screenshot enthält.

Automatisch generierte Beschreibung

# 03.01.2020, 09am to 11am, 1-to-1 with Michael Notter

* The cluster size distribution graphic can be supplemented with a logarithm transform.
* EDA on individual attributes is recommended. Examples
  + Relationship of century, pubyear, …
  + Additional, separate flag indicating that a record represents a map.
  + …
* The first PCA done with title data raises too slowly for the data being useful for information on similarity.
* Remove PCA method as an analysis step for understanding the data.
* A PCA can be done on the final features of the model with the goal to find out the prediction relevance of each feature.  
  Just an input, not to be mentioned in the proposal.
* Some thoughts have been discussed on unsupervised learning with a neural network. The idea could be to start with a PCA and transform the full data to a lower dimensional space and afterwards run the transformed data through a neural network for deduplication.

This method might also be used for the preclustering step?

* Take the proposed approach written in the proposal.
  + TSNE projection as an example. Reduction of N-dim space to 2-dim and then do clustering. Have a look at that.
* **Next steps**
  + **Submit the proposal** : The proposal will not be accepted, yet, though. Reason is that the epfl team will not be allowed to support the project work, anymore, as soon as the proposal has been accepted.  
    The proposal will be a *GO* from Michael’s side.
  + Start with data preparation until a first prototype of the pair-wise feature matrix is present. The book a 1-to-1 with Michael, again.  
    Check-in the implementation artefacts in github, in the proposal’s repository.
  + Michael will send me a 1-to-1 link, so that I can contact him directly in case of …
    - Questions,
    - Need for support,
    - Discussion of intermediate implementation results.

Book 1-to-1’s with Michael to discuss intermediate results with epfl in the course of the project implementation.