**Machine Learning Engineer Nanodegree**

**Capstone Proposal (LANL Earthquake Prediction from Kaggle Competition)**

Tobias Steidle

([tobias.steidle@softwaredev.de](mailto:tobias.steidle@softwaredev.de))

April 00th, 2019

**Proposal**

## Domain Background

Due to their devastating consequences, earthquakes and their prediction are one of the most important problems in the geosciences. The National Earthquake Information Center registers between 12000-14000 earthquakes per year. The number of fatalities varies greatly. Between 2000 and 2015, an average of 50100 people dies each year. The economic losses, which amount to billions every year, have not yet been taken into account.

Even if earthquakes cannot be prevented, a more precise prediction of when, where and how strong an earthquake will be would be of great help. This would increase the warning time and reduce the number of fatalities.

## Problem Statement

So far there are no reliable predictions for the occurrence of earthquakes. So far it has been possible to circle the location and strength of earthquakes, but not the time. After an earthquake, data records usually show small aftershocks and certain patterns. The problem is that the sensors often measure the same thing without a dangerous earthquake following. No method is so sophisticated that earthquakes such as thunderstorms can be predicted.

## Datasets and Inputs

(approx. 2-3 paragraphs)

In this section, the dataset(s) and/or input(s) being considered for the project should be thoroughly described, such as how they relate to the problem and why they should be used. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included with relevant references and citations as necessary It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.

In diesem Abschnitt sollten die für das Projekt in Betracht gezogenen Datensätze und/oder Eingaben gründlich beschrieben werden, z.B. wie sie sich auf das Problem beziehen und warum sie verwendet werden sollten. Informationen wie die Art und Weise, wie der Datensatz oder die Eingabe erhalten wird (wurde), und die Merkmale des Datensatzes oder der Eingabe sollten bei Bedarf mit relevanten Referenzen und Zitaten versehen werden. Es sollte klar sein, wie der/die Datensatz(e) oder die Eingabe(en) im Projekt verwendet werden und ob ihre Verwendung angesichts des Kontextes des Problems angemessen ist.

## Solution Statement

There are several possible options to solve the problem. It is possible to solve approaches with a "classical" ML-algorithm like Random Forrest as well as with a neural net. Since the seismic signals are similar to audio signals, methods of audio signal analysis can also be used. First, an MFCC (Mel Frequency Cepstral Coefficients) is used to check whether some MFCC values have a linear relationship to time\_to\_failure. These values can be used for training. The predictions determined are then executed and evaluated on the test data set.

## Benchmark Model

As this is a Kaggle Competition, the Public Leaderboard is a good benchmark for the planned solution. Due to the restriction of a maximum of 2 submissions per day, part of the training data can be used as test data to check the effectiveness of the solution. My personal goal would be to be among the top 20% on the leaderboard. Currently, the best value is a Mean Absolute Error (lower is better) of 1.303. The value on the Leaderboard is calculated with approximately 13% of the test data.

## Evaluation Metrics

The used metric is the Mean Absolute Error. The Mean Absolute Error measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

= Number of predicted values  
 = real value  
 = predicted value

For each record oft he test dataset the prediction error is calculated. Convert each error to a positive figure by taking the absolute value for each error. Finally, calculate the mean value for all recorded absolute errors.

## Project Design

Daten analysieren

Visualisien der Daten

Noise hinzufügen  
Verschiedene Algorithmen testen

Keras  
Scikit  
Python, Jupyter Notebooks, Panda, Matplotlib and other if necessary.

(approx. 1 page)

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.

Fassen Sie in diesem letzten Abschnitt einen theoretischen Arbeitsablauf zusammen, um eine Lösung für das Problem zu finden. Besprechen Sie gründlich, welche Strategien Sie in Betracht ziehen, welche Analyse der Daten vor der Verwendung erforderlich sein könnte oder welche Algorithmen für Ihre Implementierung in Betracht gezogen werden. Der Workflow und die Diskussion, die Sie anbieten, sollten sich an den Qualitäten der vorherigen Abschnitte orientieren. Zusätzlich wird empfohlen, kleine Visualisierungen, Pseudocode oder Diagramme einzubinden, um das Projektdesign zu beschreiben, aber es ist nicht erforderlich. Die Diskussion sollte Ihren geplanten Arbeitsablauf des Capstone-Projekts klar umreißen.

Before submitting your proposal, ask yourself. . .

Does the proposal you have written follow a well-organized structure similar to that of the project template?

Is each section (particularly Solution Statement and Project Design) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?

Would the intended audience of your project be able to understand your proposal?

Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?

Are all the resources used for this project correctly cited and referenced?

Bevor Sie Ihren Vorschlag einreichen, fragen Sie sich selbst. . .

Folgt der von Ihnen geschriebene Vorschlag einer gut organisierten Struktur ähnlich der der Projektvorlage?

Ist jeder Abschnitt (insbesondere Solution Statement und Project Design) klar, prägnant und spezifisch geschrieben? Gibt es mehrdeutige Begriffe oder Phrasen, die einer Klärung bedürfen?

Wäre die Zielgruppe Ihres Projekts in der Lage, Ihren Vorschlag zu verstehen?

Haben Sie Ihren Vorschlag richtig Korrektur gelesen, um sicherzustellen, dass es nur minimale Grammatik- und Rechtschreibfehler gibt?

Sind alle für dieses Projekt verwendeten Ressourcen korrekt zitiert und referenziert?

*[1]* [*https://www.iris.edu/hq/inclass/fact-sheet/how\_often\_do\_earthquakes\_occur*](https://www.iris.edu/hq/inclass/fact-sheet/how_often_do_earthquakes_occur) *[2]* [*https://www.statista.com/statistics/263108/global-death-toll-due-to-earthquakes-since-2000/*](https://www.statista.com/statistics/263108/global-death-toll-due-to-earthquakes-since-2000/) *[3]* [*https://www.kaggle.com/c/LANL-Earthquake-Prediction/*](https://www.kaggle.com/c/LANL-Earthquake-Prediction/) *[4]* [*https://www.kaggle.com/c/LANL-Earthquake-Prediction/data*](https://www.kaggle.com/c/LANL-Earthquake-Prediction/data) *[5]* <https://en.wikipedia.org/wiki/Mel-frequency_cepstrum>  
[6] <https://www.kaggle.com/c/LANL-Earthquake-Prediction/leaderboard>