Project name: KAGGLE-VOLCANOES

Team members: Maiken Veedla, Agnes Peetrik

GitHub repository: https://github.com/maiken-veedla/D2-KAGGLE-VOLCANOES

Task 2. Business understanding

- Identifying business goals
    - The project is about predicting volcanic eruptions like today scientists are predicting the weather. Even one foreseen volcanic eruption means that people living in that area can evacuate in time and that possibly saves thousands of lives. Eruptions are challenging to predict because the patterns of seismicity are different for types of volcanoes and therefore difficult to interpret. For very active volcanoes scientists can only predict eruptions few minutes in advance and that is not enough for saving lives.

      Just in 2019 there were 73 confirmed eruptions, currently there are many volcanoes with continuing eruptions and each day there could be a new one without a warning.
    - Our project is not meant for a certain business, but rather is meant for societies living in such areas that possibly could be affected by volcanic eruptions. Especially for these communities that live near (very) active volcanoes.
    - We consider our project to be successfull if we are in the better 50% of the competitors of the competiton.
- Assessing your situation
    - For resources we have the data given in the Kaggle competition. In addition to that we have a chance to discuss with other competitors and also ask from the authors of the Kaggle competition from the National Institute of Geophysics and Volcanology.
    - The Kaggle competition that we are competing in with our project is ending on 30th of December 2020, but because our course has it's own deadline before that then the competitions actual deadline is not valid for us. It is said in the competition rules that we can use the given data for any purpose and also use other external data from the Internet for the competition that is available to all participants.

    - We did not find any risks or continqencies that could cause a delay the completion.
    - Terminology that is relevant to our project:
      volcanic eruptions - occurs when magma is released from a volcano;
      magma - a hot fluid or semi-fluid material below or within the earth's crust from which lava and other igneous rock is formed on cooling;
      seismic signals/waveforms/activity - an elastic wave in the earth produced by an earthquake or other means;
      seismic sensors - an instrument to measure the ground motion when it is shaken by a perturbation;
      active volcano - a volcano that has had at least one eruption during the past 10,000 years;
      dormant volcano - an active volcano that is not erupting, but supposed to erupt again.
- Defining your data-mining goals
    - Our goal is to predict volcanoes time until eruption based on given seismic data around that volcano.

Task 3. Data understanding

- Gathering data
    - Our data has been gathered by Italy's National Institute of Geophysics and Volcanology and made publicly available on the Kaggle website as part of a predicting competition.
- Describing data
    - The given data has already been separated into training and testing data by the creators of the competition. The *train.csv* file has the metadata for the training files – an ID code for the data segment, that matches the name of the associated data file, and the target value, the time until the next eruption. All the data files, both training and testing, contain ten minutes of logs from ten different sensors arrayed around a volcano. The readings have already been normalized within each segment, to ensure that the readings fall within the range of int16 values.

This given data is all we can use to complete our data-mining goals. Gathering more data or identifying the exact sensors may be possible, but this would ruin the competition.

- Exploring data
  - We have data files on 8951 volcanoes, 4520 of these files are for testing and 4431 files are for training. Every data file has around 60000 rows of seismic data.

    Mostly exploring the training data is important in the beginning, because that is what we will use to construct the prediction model. Starting with given time to eruption to the training data files. The median value for this is 47125, but the minimum value is 25 and the maximum value is 1770479433. Most of the values are in the thousands, only twelve values are under 200 and about 170 values are in the millions.

    The data files for every volcanoe have 10 attributes (all seismic sensors) that range from negative 10000 to positive 10 000.
- Verifying data quality
  - This data has been carefully chosen specifically for this competition by the scientists of National Institute of Geophysics and Volcanology.

Task 4. Planning the project

Task	Maiken	Agnes
Finding a topic and setting goals	2	2
Preparation for the introduction	1	1
Presenting in the introduction session	1	-
Analysing the data	5	5
Data to Pandas datafram in Jupyter notebook	-	1
Visualizing and plotting the data	3	3
Predicting the time to eruption (using diferent models)	10	10
Summarizing	3	3
Making the poster	5	5
Total	30/30	30/30