Final Project Proposal

Group Members: Leon Kriesmair, Victoria Vorre, William von Tucher

1. Prediction Problem

We want to predict the size of wildfires in the US to help with the triage of discovered fires using vegetational data, weather data and a measure of remoteness. We will either try to predict fire size numerically (Regression approach) or try to predict the fire size class (Classification approach). In the second case binary classification (catastrophic or not) or a multi-class classification is possible.

2. Dataset

The dataset we use is a sub-sample of the Fire Program Analysis fire-occurrence database (FPA FOD) found on Kaggle. It includes a random sampling of 50,000 fire samples of originally 1.88 Million US Fires. The dataset was combined with historical weather data at specific lat/long coordinates, historical vegetation data and a metric representing the measure of the remoteness of a fire.

3. Approach

Since we are dealing with missing values, we first have to deal with those observations, either eliminating rows or replacing the data. Then we decide on what features we want to use in our prediction models which might be tied to available data. Looking at some distributional measures we will decide which kind of problem we will solve (regression / binary classification / multi-class classification).
Final model decision depends on the previous step but we want to use a blend of simpler and more complex models. Considering our presentation on tree-based methods we might lay a slight focus on those, comparing popular algorithms like LightGBM or XGBoost with more naive methods. But we will also build a deep neural network as Leon is taking the course *Deep Learning* currently.

4. Planned Division of Work

We will most likely divide the work as follows:
As a group we will do EDA and clean the dataset and decide how we deal with the missing values together to gain a good overview of our dataset. Together we also decide on the target variable design (regression vs classification). We will then continue to build our different predictive models. Everyone will at least build one predictive model on their own and on regular occasions present their progress to the group. Leon will focus on the neural network. William and Victoria will focus on the tree-based methods. Depending on the final target design the final model will be designated according to previous workload. Even though each member will focus on their model it is important for every member to understand each model. Both for the presentation and to check for mistakes. Therefore we will emphasise both group as well as individual work. Analog to the models each member will describe their models in the final report, as well as in the presentation. In the final report as well as the Python code we will indicate which parts have been primarily written by each member.