The current state of your project: what works, and what doesn't. This needs to match the code and data committed in your team's repo on GitHub.

The random forests model runs without any errors, and we can print out the predictions and labels. We believe the high losses are due to the data itself, and are looking for ways to reduce the loss. Refer to the challenges section for our detailed plan.

The multiple regression model runs without any errors. The model incurs a high loss and this challenge is mentioned in more detail below.

Any feature changes to the proposal. As we discussed in class, no reason is necessary for M1, but any feature changes in M2 et seq will need a technical justification. Obviously, feature changes need to keep in spirit with the project, so you can't for example remove all the hard parts :)

Nothing major has been changed for pair team 1. We are currently facing some challenges, as outlined below. We may need to choose a different output of our model, but these changes will not alter the spirit of the project.

Pair team 2 researched and determined that RNN was more suitable for predictive sequencing of text and less for prediction for a regression task. We then decided to explore a Multiple Regression Model to predict the size of a fire given multiple inputs.

The current challenges / bottlenecks you are facing, both technical and otherwise, and what you are thinking of doing about it.

The loss for the random forest model is somewhat high. Here are the tests pair team 1 plans to run:

1. remove outliers
2. have dropout values
3. use only the randomly generated ‘no fire’ data - not the ones with 0 hectares burnt (as clarified by piazza post, firefighting teams may have intervened quickly. Hence the climate data may not actually reflect whether or not a fire developed).
4. Change the labels (to be a classification task). In this case, we would report the probability of a wildfire developing, rather than using fire with size <= 0.1 as “no fire”. We would not be able to report the size of wildfires, but changing the problem to a classification problem is an option to look at.

The Multiple Regression Model gives a high loss as it deviates from the expected result. Replacing the missing data with the mean of the column could have caused the issue or a poor correlation between the inputs and output. We are thinking of figuring out an accurate method to fill in the missing data, build a new model that can handle missing data well, or use some strategies mentioned above.

For each team member, what tasks were done and which tasks are underway.

Student 1 (Anna Wang): Did half of pair task 1 (pair programming). Went to office hours to get help on loss functions. Will investigate/test regularization techniques for random forests for milestone 3.

Student 2 (James Ardian): Did half of pair task 1 (pair programming). Went to office hours to get help on loss functions. Will investigate/test normalizing features for random forests for milestone 3.

Student 3 (Michele Mai) collaborated with Student 4 to research the applicability of the RNN model to our project. We realized RNN was more suitable for predictive sequencing of text and less for our project. We then decided to explore a Multiple Regression Model to predict the size of a fire given multiple inputs such as temperature and wind direction. Finally, Student 4 computed the loss between the expected and predicted results. Did pair programming with Student 4. Currently in the process of investigating a way to minimize the loss of our model.

Student 4 (Alex Liu) Worked alongside Student 3 (Michele Mai) to attempt to create a RNN model for our project. In addition, we worked together to create a multiple regression model for the dataset, but found that it returned poor results.