Contents

Plan, Cost/Benefit Analysis and Ethical/Legal Considerations Report

K Bacon, N Higgins, P Kim, B Townsend

November 8, 2017

Abstract

Ensemble models are a key method for tackling uncertainty in the way that we model the earths climate. They involve running many simulations each with different parameters, with the variance in these parameters representing the known uncertainty in measurements such as the initial conditions. In certain cases some of these simulation outputs should be discarded as outliers due to the unphysical nature of the results, but often these unphysical outputs are too subtle for somebody not expert in the particular field to confidently evaluate. This project aims to produce a solution to this issue by utilising machine learning technologies to understand patterns in the opinions of experts, allowing their careers worth of knowledge and experience to be easily distilled into a piece of software that is available to anyone who requires it. Additionally, as ensemble sizes, that are currently limited by the availability of computational resources, continue to grow larger and larger it becomes progressively less feasible to manually evaluate each model. This software should be able to assist in this case as well. The initial focus of this project is to develop a user friendly web tool capable of learning the opinions of individual vegetation experts whilst abstracting any of the machine learning to a level that can be understood by anybody, independent of their level of knowledge , N Higgins,in this field.

We certify that all material in this document which is not our own work has been identified.

1 Problem Definition

Climate research is a field concerned with understanding how the natural systems come together to form our climate. It is an expansive field that is comprised of many different disciplines, with many different potential applications for its insight. A majority are concerned with our climate's rapid rate of decline over the past few decades, focusing on publicising our impacts and pointing out the natural systems we have effected. Others are using the information more directly to help communities plan their agricultural layouts, warning of potential droughts, heavy rainfall, and planning for extreme weather systems like tornadoes.

The world's climate is predicted through an aggregated system of localised models known as a perturbed physics ensemble (PPE), which has been the product of years of engineering. An Ensemble forcast is a numerical climate prediction method which is a form of Monte Carlo analysis. The ensemble utilises a set of forecasts to reason on a number of possible future states of our climate. This is in an effort to aleviate three primary forms of uncertainty in forecast models: The errors introduced by the use of imperfect initial conditions, parameters which become erratic due to the chaotic nature of our climate; and inaccuracies programmaticaly introduced into the forecast, differences of values overtime become noticeable as a consequence of approximations of mathematical equations; and finally unknown relationships between particular parameter values as they evolve through the simulation. The ensemble aims to provide a spread of possible eventualities, and the current climate state is meant to fall within it. It is this spread of future possibilities that used to indicate the level of uncertainty the system has.

As the information about the state of our environment becomes more accessible, the conversations humanity are having about the correct course of actions becomes ever more inclusive. The accuracy/uncertainy of the ensemble has been steadily improving but is seen to provide insufficient evidence when suggesting areas of focus for policy makers or activists. Some systems like fluid dynamics have been modelled to such a level of specification that any imaginable improvement would be negligible, However, many sections of the ensemble are still rather unproven. As Ensemble sizes have grown, access to computation resources have also increased and so too has the push to increase the forecasting ability of the system.

Furthermore, not all models that are created agree on the trends and the impacts of those systems and have a large variance concerning rather critical aspects like temperature. This discrepancy is a focus point of some to imply a lack of understanding of the subject, but this is likely due to subtle differences in the multiple methods researchers have used to simplify our natural world.

These localised models can be vastly different in their respective level of specificity as they are generally handled by, teams of researchers. They all aim to mimic some underlying trends innate to our habitat and collectively illustrate their effect. It is a requirement that these systems are used in collaboration as it better reflects our uncertainty of their impacts. Identifying an individuals system's outputs is rather difficult, as these systems are typically so large that they cannot be isolated. Additionally, historical climate information collected to date is unable to identify the degree of impact each system has had, the cause is largely speculative. That being said, some climate systems are known to cause cascades of interactions which typically lead to feedback loops of cause and effect. These interactions tightly lock systems together enforcing the current framework of collective analysis.

Assessment of any particular aspect of the aggregated system is measured only via the impact it has on the overall outcome, and despite the method being a generally useful indication of success, it does fail to illustrate unexpected behaviour in the PPE. A change that is made to a region might interact negatively with another section causing it to behave erratically. This behaviour means that the change results in an overall accuracy decrease regardless of whether the change genuinely improves readings from its target region. The erratic behaviour is primarily due to a lack of understanding of the system in entirety, and inconsistencies in our own knowledge. This is considered systematic of a PPE, as the precise values of the system parameters are not known. Through this method, the new addition is rejected as its error is misunderstood.

In an attempt to navigate around this problem, researchers typically allow for the accuracy to decrease by comparing the accuracy of a change with a collection of new experimental changes. The changes that are most promising (and not necessarily better than the current accepted model) are chosen to undergo further research. This ensures that development is not driven by past mistakes and promotes innovation across the platform.

At this stage of the process, researchers have identified a glaring problem that may potentially be resolved through an engineered software product. The task of deciding whether the changes could be considered useful or an improvement becomes a discussion point and requires a group of researchers

to have an intellectual debate. For a given cycle of changes, they must compare the merits of each model and order them respective to their intrinsic value, which is rather time-consuming. Additionally, as researchers have a limited amount of time to offer, only a limited number of models can ever be discussed. In a preprocessing step, a collection of rules are used to create a subset of viable models, a set that is not proven to be optimal.

Our project is tasked with resolving these issues by automating the decision process. With an automated system for ranking model worth, researchers would be able to complete entire cycles faster, identify internalised biases in model output and work more effectively with a larger group of invested parties. We have been tasked with producing a system that can learn the intricate focus points and values of an expert such that we might be able to predict their opinion on a model without requiring them to be present. This would allow a single individual to produce collections of outputs for various climate models and evaluate them at their leisure, gaining the insight and guidance of an entire community of scientific researchers, at the click of a button.

2 Solution specification

It is the intention of this project to: improve communication between researchers and the public; to facilitate the development of innovative improvements on the current climate model, and to increase inclusiveness and accessibility in the field. To achieve this we have identified that our solution should be formed from two distinct but coupled code bases: A web server and a machine learning based python package.

A web server architecture provides flexibility and utility in being the graphical interface for the machine learning package while providing a useful level of abstraction for non-computer inclined users. The web application facilitates the basic use of the package while providing a platform for more complex operations to be formed from the API the package presents. Interactions with the application would be similar to that of any other popular web service, ensuring a minimal barrier to entry and allowing for scalability and connectively with the community.

The machine learning package contains the code responsible for the intelligent representation of the expert, it offers the means for generating and managing multiple expert models that can be used to evaluate a new climate models outputs. Compiling the core fundamental system within a package allows it to have many options in its future implementation. The package can be verified and tested separately from any system so that users can be confident in its capability. Presenting a simplistic API, the package can quickly be built into an existing system or new one and begin providing useful insight. As the project is likely to be fluid in design, changing with the needs of the client, the system's portability is an important aspect.

Throughout the development of the project, functionality in both the web server and the machine learning package is required to better analyse the problems faced. At this stage it is unclear what machine learning techniques will provide the best results, and additionally what kind of analytical insight the experts would appreciate most. Work in this pursuit does not form part of the end solutions but is a vital intermediary step for its identification.

2.1 Web server

The web server is the primary visual aid to the project. It's requirements are based around the visualisation of information and efficiency of transfering insight. It will be responsible for interacting and operating the machine learning package but is a complementary piece.

2.1.1 Model Uploading

For use of the system, vetted climate model outputs (CMO) are required to form the feature information for the expert models. It is a requirement of the solution owner that there be a protected page dedicated to uploading these models. Furthermore, only the solution owner and vetted experts should have the permissions necessary to perform uploads. Access to the rest of the user-base should be restricted.

Alternatively, to not restrict the potential for available data, all users might be able to upload CMOs. In this instance, the application would need to validate the models themselves, ensuring that malicious or duplicate information isn't recorded. The accuracy of the model would not come into question, as even poor CMOs provide useful training data.

2.1.2 Labelling

Individual to a particular expert, the page presents a visualisation of an unseen CMO and question pairing. The intention is that the expert filters through the outputs to score how well the said output answers/resolves the question. The page facilitates in the filtering of the content and contains a form to submit the output's score. Each submission dynamically updates the page with a new pairing for continuous labelling.

The site is required to immediately upload the labelling of a model output to ensure that no labelled information is lost. Experts are envisioned to visit the site erratically, so if they decide to perform some labelling, there should not be any time obligation enforced upon them.

The CMOs are required to be visualised as their information is difficult to interpret natively. The user requires a visualisation that can correctly map the CMO onto a projection of Earth, in a way that helps illustrate locations of high/low concentration for the given output. Equally the user requires a method of shifting through the outputs without the page becoming cluttered with information.

Information that does not help answer the question posed is not required and clutters the expert's view. The user requires the ability to filter out climate model information that is unnecessary. Additionally some information is best utilised when in contrast to other outputs, the user requires the ability to select multiple outputs for a quick comparison, or to merge the output, as to help them in their understanding of the model effectiveness. The minimum requirement is that the information is displayed with a visualisation and the expert can navigate the output.

2.1.3 Evaluation

The evaluation aspect of the application is open to all stakeholders. the Web server should contain pages that allow expert models to evaluate collections of CMOs. The result should be displayed using in an intuitive manner, providing easily extractable information.

The user requires a means of selecting which experts will conduct the evaluation. A minimum requirement would be that the user is able to select one, ideally they would be able to select a collection. The page is required to present information regarding the availability of expert models, the model's completeness, and the natural preferences of the user for quick selection.

The user requires functionality to help upload CMOs to the system. Systems that allow for drag-an-drop functionality and file explore interaction would be optimal, but any method would require the ability to submit a collection of files at once. An alternative solution would come in the form of multiple singular uploads with delayed evaluation, however this would be cumbersome to the user.

The user requires the evaluation process to end with a ranked visual representation of the quality of the climate models uploaded. The consensus of the expert community should always be shown for reference, but the selected experts should have their values listed, and a means of filtering the information should be implemented for insight to be taken.

2.1.4 Account management

A dashboard for a particular expert providing analytical information on their interactions with the system and the state of the expert model that represents them. Accessible to all, it is a method of illustrating uncertainty and validity in an expert.

An expert should be able to see the expected performance of their representation at any time, and this information should be updated as soon as training information has been provided. This information should also accompany the evaluation process indicating to the user the uncertainty this particular expert would have.

Analytical information collected as the expert navigates the site should also be present, along with their configuration of the labelling environment, to highlight potential biases or regions of disinterest. This information should be visualised alongside the representation to indicate areas where focus should be given to for a complete model.

Stakeholders should be able to navigate to this page to have a more detailed view of the expert's metrics, and possibly in future iterations allow for communication. They should not be able to see any individual label information, nor any analytical information pertaining to bias. It should provide only non-private information.

2.2 Machine Learning Package

The machine learning package is standalone package and can be intergated with an arbitrary system.

2.2.1 Training

There is a general requirement for the system to learn/be taught the innate trends in opinion the expert holds, but there is no requirement for a particular method. The process by which this is achieved shall not be known until analysis of available techniques and their viability is complete. An optimal solution would be a technique that can undergo on-line learning and be inexpensive. On-line learning complements the slow-paced nature of data collection and allows for quick inclusion into the expert model. The inexpensive application of the model would go a long way in reducing the load on the server and ensure quick evaluation of CMOs.

Alternatively it would be possible to use a technique not capable of on-line learning, however, it would require re-creation whenever new labelled information is presented. This would result in a noticeable downtime of a particular experts model on the server, or would require a change to the requirement of immediate teaching.

The teaching process is required to be abstracted away from the user of the package. The parameters and the techniques for creating a model should be identified during our analysis phase and become a static aspect of the model. If parameters are to be dependant on the amount of available training data, their variation would be hidden from the user.

2.2.2 Management

It is a requirement that the expert models persist outside the operation of the system and be transferable. Due to this, the package must be responsible for the generation of serialised model objects. More to the point, the package must be able to manage the interaction of any particular model with a system in such a way as to not reduce or affect service. Namely a system that simultaneously receives commands to use a model in an evaluation, and to train on a point, should be handled appropriately.

2.2.3 API

It is a requirement of the package that interaction is as simplistic as possible. A single API should be responsible for interactions between the web server and any expert model. Functionality should exist within the API to manipulate models and extract analytical information from them. Error metrics and accuracy scores are required to validate the model as well as indicate uncertainty.

3 Maintenance Plan

This section will cover what measures will be taken to ensure that the codebase is well maintained throughout the project and after the project is completed. Maintaining the codebase entails ensuring that all its features remain up to date and remain in the scope of the project and also ensures that the addition of new features does not break other features or render them obsolete.

An agile approach to development ensures good maintenance of the project on the macro scale. All new features are discussed and directly assessed by the problem owner so that they are guaranteed to be within the scope of the project, and the scope itself is also regularly revised. On top of this, the developers are split to manage different features of the project, therefore, ensuring that all features work together in a cohesive fashion. On the micro scale, frequent testing using automated and manual methods further ensures the code is well maintained during the project. A Bitbucket repository will allow individuals within the group to upload new code, as well as keeping existing code up to date. Not only does Bitbucket create a secure site where loss of work is not a concern, and it allows for rollbacks in case of errors. Bitbucket makes collaborative development particularly easy through the means of feature branches and pull requests. Our system will be implemented and tested on a University supplied server.

Once the project is completed, the required long-term maintenance is low. The learner will continue to improve using automated online learning and its performance can be monitored and validated by expert opinion to ensure that the learner is still accurate. The architecture of the webtool and back-end system is designed to be as flexible as possible when it comes to updates or repairs by avoiding close coupling of the system and planning for the future addition of new features and scalability. Regular

checks for bugs in the webtool and monitoring the performance of the server throughout automated checks is enough to maintain the code.

4 Cost/Benefit Analysis

As this is a research focused project, there is no plan to sell or monetise the intellectual property of the codebase at the end of the project timeline. Rather, this project adheres to a more open source approach. The webtool and all its functionality will be free to access and use. Time of the experts is of high value to the project as their insight is crucial to the development of the system. Their primary responsibility however is to the work they are already undergoing. All efforts should be to minimise our time requirement for them, and we should aim to fully utilise the time to collect relevant information. With this project we hope to accurately represent the opinions of these experts. With this representation the work that would take the real experts hours even days would be done in mere seconds. On top of this, the knowledge of these experts can now be easily accessible to the world.

We require the use of a university server to be able to host our webtool and perform the expert model training. There is a large computational expense to training machine learning techniques and as we will be experimenting with multiple methods while determining the most appropriate method. Due to this, it is foreseeable that training will have to occur many times. The server will perform the training.

4.1 Development Costs

A significant initial cost to this project and most other machine learning based projects is the time taken to pre-process an appropriate data set for training and testing of the learner. For this project, a supervised learning approach is used therefore training and testing of the model requires labelled data. In the case of this project, the data (output from surface vegetation models) must be given a score by experts for multiple questions therefore, each vegetation model must be labelled multiple times. When training a machine learner, the larger the training (and testing) data set, the more reliable the learner predictions will be. This initial cost can be very large, but the larger the cost, the better the pay-off is. This cost can be partly offset by incrementally adding to the labelled data set as development progresses. To begin with, vegetation model outputs can be labelled once a prototype webtool is set up and used for preliminary testing of the prototype learner with the aim of increasing the size of the labelled set as much and as soon as possible.

This project follows an agile approach to development therefore frequent meetings with the project owner are necessary throughout the development phase of the project. Meetings with the project owner happen on a weekly basis, creating an additional cost with regards to the project owners time. An agile approach to development also means that testing and maintenance of the codebase happens simultaneously to development. And so the pre-completion maintenance costs of the codebase must be taken into account. The work involved in the pre-completion maintenance of the codebase is described in the aforementioned maintenance plan. The cost of testing the codebase for this project factors in solely for user acceptance tests (UATs) as these require the project owner themselves to test an iteration of the product.

A dedicated server is required since the only interface between the user and the learner is the webtool. The server is required during the development phase so that development can take place in the appropriate environment and that reliable tests can be carried out. During development, the webtool will be hosted on university provided server space, therefore, the server cost is the resources taken away from other potential uses.

The final significant cost to be considered is the time devoted by us, the developers, towards the completion of this project. Time spent on this project is time spent away from other potential projects.

4.2 Post-Completion Costs

After the project comes to a close, the server, data, and code will be given over to Dr. F Hugo Lambert/the University. He and whomever he chooses to work with will then be responsible for the further development and maintenance of the code, data, and site.

Once the development phase is complete, there are still maintenance costs to consider. General upkeep of the webtool and planned or unplanned patches and improvements. Once the webtool is available to the wider scientific community there is the inevitability of required bug fixes to account for as well as

planning for performance checks to ensure the learner is appropriately adjusting to new labelled datasets and making accurate predictions on real data.

The team has a future ambition to intergrate the system into a docker container. We will develop the project with docker in mind however due to server restrictions we will not be implementing it currently. This will ensure longevity and ensure portability along with easy of deployment. Docker facilitates packaging the codebase along with all dependancies resulting in a system that can run anywhere and is agnostic to the host platform.

The ability of the learner to carry out online learning means that training can continue post-completion. However this means that experts time is still required to label new data so that the prediction of the learner can be improved. However this cost has less impact compared to the development phase since the learner is already trained to an acceptable standard, therefore, any further training carried out post development, although still important has less impact on the performance of the learner.

4.3   Benefits

The goal of this project is to provide an efficient way of choosing the best models using a large selection of model outputs. The model outputs represent the observed effects of surface vegetation on carbon feedbacks over the last 200 years. These model outputs are relatively cheap to obtain as they are based on existing observational data hence they can be run a large number of times. However running these models into the future is expensive and requires greater computational power. This computational power is expensive to use therefore only the best models should be chosen. Creating an efficient method to select these models will save the expense of using sub-optimal models for future predictions.

Furthermore, by efficiently sorting model outputs, expert time is saved from manual sorting. Hence one of the main benefits of this project is to save expert time spent manually qualifying model outputs by using a machine learner to do this sorting automatically. The long-term aim is that the expert time spent labelling data is largely outweighed by the time saved by this automation. On top of this, expert knowledge on model output analysis will be available to the public in a far more accessible way.

In an agile approach to software development, having the problem owner closely involved with the development process ensures that the scope of the project is well defined and that the codebase is always meeting the problem owners requirements. Agile development also allows for flexibility of the project, new functionality can be added or removed throughout the development without major backtracking. This is valuable for this project in particular where the final requirements have a degree of flexibility and the requirements may change during the development process. And so the cost of the project experts time is outweighed by the benefit of a more efficient and flexible development process leading to a better end result.

Using an on-line learning approach means that the learner can be improved over time without input from the developers. Re-training of the model is not necessary and so a significant portion of learner maintenance is done automatically. New labelled data sets can only be supplied by trusted experts, therefore, the continuous learning is reliable.

5   Ethical Considerations

Ethical, legal and professional considerations are important for any software engineering project. Machine learning and data-driven systems require an extra level of care due to legal aspects of handling data, as well as the ethical aspects of the decisions that are derived from it. This project has a very specific set of issues that should be dealt with as the project itself is designed to simulate the professional opinions of experts, something that has both the potential for abuse and consequently potentially damaging consequences for the professional.

5.1   Project Specific

In this subsection, issues that relate directly to this project will be coupled with proposed courses of action to give a full picture of the issues and the decisions that the development team will make to overcome them.

5.1.1   Ethical

The core ethical issues of this project fall into four categories.

- Ethical issues relating to the misrepresentation of academic experts opinions and the knock-on effects of this on their careers.

- The potential for this system to be used maliciously.

- The potential for inaccuracies in the system to affect scientific results or influence legislation.

- Holding of personal information.

Misrepresentation Of Opinions: The underlying principle of this project is to generalise a researchers opinions and preferences to estimate what their opinion is of unseen data. Whilst this can be extremely useful in time-saving these so researchers time and not having to give their opinions on every single piece of data it has the potential to drastically misrepresent the opinions of these experts. In the worst case, having their name attached to opinions that are not representative of their own may be damaging very to their career. To mitigate against this risk it is suggested that the page that produces inference results will include a disclaimer. This will clearly state that whilst the results produced during the inference step is supposed to represent their opinions it is not their opinion, and may sometimes not produce a faithful representation. As an additional safeguard at the training phase, metrics such as accuracy or correlation metrics calculated on a validation set will be shown to the expert, before he/she decides whether to publish his "opinion predictor" to be usable by the public. It is believed that these two safeguards will mitigate the risk as much as possible.

Inaccuracies in the System Affecting Scientific Results or Influence Legislation: Related to the above point is the capabilities of the system to bias scientific results and as a result potentially influence legislative decisions about climate change. The solution to this is likely the same as the above point, including disclaimers that the produced results are not necessarily 100% representative.

Malicious Usage: Climate change is a very contentious topic that has been shown to drive people to extreme lengths to fight for their side of the argument, even to the extent of spending vast amounts of money or committing crimes to convince people of their opinion [2]. As a result of this it must be ensured that the system is:

- As secure against hacking as possible and keeping thorough logs of usage to help understand what has been compromised upon a breach.

- Enforcing that accounts cannot be set up in the names of others so that they can impersonate them.

- A method of disambiguating experts with the same name, eg. Institution

- Some form of verification status. This may be as simple as verifying that account holders for the "experts" possess *.ac.*, *.gov or *.edu email addresses.

Alternatively to the above ideas, it may be possible to simply restrict the availability of account creation to invite only. As was seen in the Climategate scandal, keeping the science open to the public is beneficial in the long run as it ensures clarity that no facts are being hidden. For this reason it should be possible for any interested party to access the inference section of the site if required, it would also be beneficial to describe the methodology for the machine learning element on the website in a way that is accessible to non-machine learning experts.

5.1.2 Legal

The data held in as a result of this project is believed to be covered under the Data Protection Act 1998 and the new General Data Protection Regulation (GDPR) 2016 due to the ability to identify the individual given the information, both explicit and implied that is retained in the user accounts of the web tool. Eg. Name and Email which when combined with the knowledge of the field of research would make it very trivial to find out precisely who the information relates to. For the purposes of this section, the GDPR will be specifically focused on due to its replacement of the Data Protection Act on the 25th May 2018. The GDPR states that anyway the data will be used or processed should be transparent and clearly presented to the user. It also states that the information should be viewable by the user and have the ability to be updated should it be found to be inaccurate or deleted upon request. All of these

points have fairly trivial solutions such as clear disclaimers and a simple accounts page showing all the information stored relating to that user as well as a delete account button to cover the right to erasure [1].

Something that is a more interesting for this context is the right to explanation that states that a user has the right to explanation about decisions that are made using their data. This is interesting as in some machine learning systems it is extremely difficult, if not impossible, to explain the decision that was learned from the data. Luckily article 9(2) allows somebody to explicitly consent to forego this right. In this case this is probably very likely to be the solution that is taken it is very possible however that certain algorithms such as decision trees or KNN may offer suitably explainable algorithms and asking the user to waiver this right may not be necessary.

5.1.3 Summary of Terms and Conditions Requirements.

For somebody who is going to be using the systems as a non-expert to label climate model run:

- Consenting their awareness that the algorithm is not to be assumed to give a totally faithful representation of an experts opinion for all cases.

For experts who are labelling data must consent to the above as well as:

- Consenting for their personal information to be held and their "opinions" offered to the general public.

- Waiving their rights to explanation of how decisions are made based on their data. Note, an explanation of the algorithms used to make the decisions will be shown, but it may not be possible to show how these algorithms make the decisions they do, which appears to be in violation.

Software License Agreements   Consideration must be taken throughout the development process to use only libraries that are free to use for academic work. Each piece of software must be considered on a case by case basis based on the licensing agreement for the particular library or software. Some useful information can be found in [3].

5.1.4 Professional

A large consideration that needs to be made is on the professional responsibilities of taking on a project such as this. This project is a very high-risk endeavour whereby it is possible that there are real limitations that are baked into underlying limits of the world such as information theory and therefore no amount of clever engineering will ever be able to solve. An example of this kind of problem is that it might be the case that the machine learning aspect of performing the desired inference on this data requires so much data that attempting to represent the opinions of a single expert may be completely infeasible due to the volumes of samples that the expert might need to label.

5.2 Domain Specific

This subsection contains issues that relate the domain of machine learning and climate but do not apply directly to this project.

5.2.1 Ethical

Technology and more prominently machine learning has a potential byproduct of significantly reducing the number of skilled jobs in certain fields, either by making people's jobs easier and therefore reducing the skill required or by removing them all together. Cases of this have been seen in areas from robots replacing factory workers all the way through to uber researching autonomous vehicles gearing up to replace their human drivers. Fortunately this is a non-issue in this case due to the fact that for the people who the developed system will be "imitating", assigning rankings to climate model data is a tiny and likely unenjoyable slice of their job. The rest bite from this process will allow them to focus more on their core research and make more valuable contributions to the field.

6 Development Approach

This project will be completed using an Agile methodology. Agile has been chosen because it allows for not only changes to the time plan, but to changes in requirements/desirables of the customer. Other module projects and the generally packed lives of the collaborators to this project will no doubt cause changes to when different components are to be complete. This factor along with the evolving idea of what is wanted, and possible, correspond well with the Agile methodology.

Python has been chosen as the language to develop the machine learner in. The reasoning behind this choice is that all group members have at least some experience working with the language, and others even have experience using it for machine learning. Pythons scikit-learn module also comes with many prebuilt machine learning techniques that are easy to comprehend and use.

In addition to the basic web development tools like HTML, CSS, and Bootstrap we have decided to use Python Pylons Pyramid. Pylons is an open source Web Framework written in python. The decision to use this as the backend framework was influenced by the plan for the learner to be written in python, allowing for easy integration. Also a group member already has experience in using this framework, making it the obvious choice.

6.1 Milestones

- API(ASAP) The API must be created and agreed upon first. This is the backbone of all the programming that will be done during this project and is what will allow us to work individually without running into compatibility issues later down the line. At this point we already have a draft of the API and are working out differences in opinion, and filling gaps.

- CA1 (10/11/2017)

- Frontend Skeleton (28/11/2017) The website pages are created with little to no functionality. Just a visual representation to give us a sense of what needs to be developed for each page, alongside how the site should be navigated.

- Backend Skeleton (28/11/2017) The server for the site is up and running and a database has been created to store the relevant data that will be used in the project.

- Data visualization (15/12/2017) This will be a crucial element in the collecting of data. The models have outputs of up to 17 vegetation types, all need to be accessible visually at the same time. Without this visualization it would be very difficult for an expert to discern the good outputs from the bad outputs, and even if they could it would be too time-consuming.

- Data Collection Functionality (15/1/2018) The page where an expert gives their opinion is fully functional. This includes visual representation of the models, user input, and data capture. Completion of this milestone is instrumental in the development of the rest of the project. Currently we have no data on the opinions of experts and without data we cannot develop a machine learning technique.

- Basic Learner (30/1/2018) Once the data collection is implemented we can begin to develop our learner. At this point the technique used by the learner does not need to be permanent, but a functioning learner is needed so that other components can begin development.

- Chosen Set of Algorithms (10/2/2018) After a learner has been created we can start experimenting with the different types of machine learning algorithms. Through this experimentation we will come up with a single/small set of machine learning algorithms that users can choose from.

- Training Stage Complete (17/2/2018) At this point an expert should be able to login to the site, give their opinions on models, have those opinions trained into a learner, and use that model to provide opinions on new models. This element can begin development once the data collection and basic learner are completed.

- Upload/Result Page Complete (3/3/2018) Anyone should be able to come to the site, upload models, select an expert, and get the results from that experts learner. At this point the site is fully functional.

- Final Report (11/05/2018) Our experiences throughout the process will be aggregated into a comprehensive report describing our insight along the way and the final state of the project.

- Presentation and Demonstration (second half of term 3) We will present our findings and demonstrate its use. The aim would be to prove its viablility as a useful tool for experts.

6.2 Risk Analysis

The main risk that this project will face is time management. The entire project is on a short time span, being that it needs to be completed by May 2018. That relatively short amount of time on top of the busy lives of the group members does not leave a lot of room, when it comes to getting different components completed. Up to this point, the group has been able to meet regularly with any real hassle. This will most likely become more difficult as the term progresses and other modules start to give out more work.

Equally impeding, will be size of the data set. At present, we have no data to train a learner on. This would not be too large a problem if we could generate the data ourselves, but the groups ability to do so will be limited, considering that only Dr. F Hugo Lambert is the only expert among us. In light of this the development of the Data collection process is crucial so that we can start collecting data. With a large amount of data, there is no guarantee that the learning methods that we implement will be able to accurately represent the experts. With a lack of data, it could prove very hard to find a correlation between model ratings and expert opinion.

Other risks to the project include lack of communication outside of term time, differences in opinion among group members, change in requirements late in the development process. The largest of these being the lack of communication outside of term time. The group members all live in different areas, there are even members who reside outside of the U.K. and time zones may make communication sporadic.

For fall back plans the first thing to go will be the unnecessary components, such as output on the features that the learner believes the expert to find most important. Further fallback could end up with the use of only one learner as opposed to a set.

References

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L119:1–88, May 2016.

[2] B. Nerlich. 'climategate': paradoxical metaphors and political paralysis. *Environmental Values*, pages 419–442, 2010.

[3] U. of Brighton. A guide to the ethical and legal use of software for staff and students. https://staff.brighton.ac.uk/is/PublishedJuly 2015.