Technology 1: Machine Learning Libraries/Platforms to Train the Model

1. Option1: SciKit

Option2: Amazon Machine Learning

Option3: Pylearn2

b. Goals for use in design

Our Algorithm should be as accurate as possible. It should be highly scalable considering we may have a lot of basketball match data. We also want full control of our algorithm, i.e. have access to the source code if we are using some library in case we may need to modify them for our project.

c. Criteria being evaluated (e.g., cost, availability, speed, security, etc)

The cost is definitely one important aspect to be considered in our project. We want to make our model available to most people so they can learn that machine learning is a power tool to use in many areas. Providing access to as many people as possible at a low rate or free is our goal.

The availability is also considered. Some service may only be available in some countries. We should try to avoid those that are only accessible in a small area.

Speed is also an important aspect of our algorithm. Making our algorithm efficient really enhance the user experience a lot.

Finally security is something to be considered in the process but not too important since we don’t have private user information. But keep in mind that a major flow in our project could cause danger to the system.

d. Table comparing option 1, option 2, option 3 based on criteria

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Technology | Economy | Security | Availability | Speed | Notes |
| SciKit | Extremely economical | secured | Open Source Library | Very Fast | The most popular Open Source Machine Learning Library on Python |
| Amazon Machine Learning | Economic | secured | Available globally based on Amazon’s server | Very Fast | Easy to start with but may don’t have all the access to the detailed algorithm. Have customer support |
| Pylearn2 | Extremely economical | secured | Open Source Library | Very Fast | Popular open source library. A large community of developers on Github |

E. Disscussion

All three options are very reliable. Amazon Machine Learning(AML) is a fast developing platform that provides easy Machine Learning model to many customers with different background to start with. The advantage of Amazon Machine Learning model is that it has great GUI that makes the process easier. And according to its introduction, AML is very closed to the purpose of our model. With data training the model, it will give a prediction based on the data. Users can also narrow down the searching areas. But because AML is not open sourced I will first dismiss this option. Our project is likely to modify a lot of algorithm in order to better meet the potentially changing needs. SciKit is currently the most popular open source library on Python and has active developers developing them and fixing bugs. While Pylearn2 is also a famous and a widely used library on Python, there is no developers responsible for developing them. Although Pylearn2 claims that they will continue reviewing pull request on Github, there is no active developer for the project.

F. Selection of best option based on criteria

SciLearn is the best option for the need of our project. It is open-sourced, reliable and has active developers developing it. It is also widely popular which means we will be more likely to get community support.

Technology 2: Server for computational power and hosting our database and

1. dOption1: Amazon Web Service

Option2: Google Cloud Platform

Option3: Oregon State University Student Engineering Server

1. Goals for use in design

Our project is likely to require strong computational power because our prediction model needs to evaluate a large amount of data. And if we want to access our model from different devices and different locations, it is important that we train our model on the cloud. A reliable database server on the cloud is also needed for the users to easily use our model worldwide and cross-platform.

1. Criteria being evaluated (e.g., cost, availability, speed, security, etc)

Choosing the server for our project should consider the cost, availability, speed, reliability. All of these factors are extremely important. The reliability is the most important factor of them. Our data is precious and the machine-learning-algorithm-trained model is peculiar. Thus we can suffer any data loss. The speed is important as well. The AlphaGo from DeepMind trains itself by playing GO with itself millions of times every day. With fast speed, our model could be more accurate. Availability is also to be considered. For example, Google Cloud Server is not accessible while OSU server and Amazon Web Service are accessible in China.

1. Table comparing option 1, option 2, option 3 based on criteria

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Technology | Economy | Reliability | Availability | Speed | Notes |
| Amazon Web Service | economical | Very secured and reliable. | Available worldwide and cross platform | Very Fast | A very popular service by Amazon |
| Google  Cloud Platform | Economic | Very secured and reliable. | Available in most countries and cross platform | Very Fast | A very popular service by Google. But GCP is not able to access in China |
| OSU Student Engineering Server | Extremely economical | Secured and reliable. | Available worldwide and cross platform | Fast on campus, moderate outside campus. | Free to OSU students, easy and familiar to use. Have support directly on campus. |

1. Discussion

When considering the server we use to host our model and data, it is extremely important to consider the factors mentioned above carefully. OSU Student Engineering Server is functionally complete. You can host Linux project there and OSU also provides database server. It is easy and familiar to use. If we have difficulty, we can directly access to tech support on campus. But it is not very fast outside campus. And OSU engineering server is not very flexible if Windows server is to be used. Also, student usually only have access to MySQL database. Google Cloud Platform and Amazon Web Service are very similar. They both are hosted by giant companies in the US. The major difference between them when considering starting a new project is that Google is not accessible in China.

1. Selection of best option based on criteria

Overall, after carefully comparing important aspects of them, I think Amazon Web Service is the most reliable and appropriate solution for our project.

Technology 3: Statistics Library to determine the importance of each factor

A Option1: [py-statistics](https://github.com/digitalemagine/py-statistics)

Option2: [scipy.stats](https://docs.scipy.org/doc/scipy/reference/stats.html#module-scipy.stats)

Option3: Pandas

B. Goals for use in design

An important aspect of our project is a statistical model to determine the importance of data. With good python library and combination of machine learning training, we will get a pretty good estimate of how these data factors affect our result.

C. Criteria being evaluated (e.g., cost, availability, speed, security, etc)

The cost are all free for the three libraries. The availability is important to be considered. The speed is important to consider when choosing the libraries. We are going to use the functions a couple times, even likely in a loop that cause higher complexity. So the basic running time of functions in these libraries are important. Security is not too important since we don’t collect sensitive data from user but should be considered for the security and reliability of the system.

D. Table comparing option 1, option 2, option 3 based on criteria

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Technology | Economy | Reliability | Availability | Speed | Notes |
| py-statistics | Extremely economical | Very secured and reliable. | Open Source Library | Very Fast | Official statistics of Python |
| scipy.stats | Extremely economical | Very secured and reliable. | Open Source Library | Very Fast | Provides more scientific functions. |
| Pandas | Extremely economical | Very secured and reliable. | Open Source Library | Very Fast | Provides data analysis support. |

E. Disscussion

Py-statistics is a powerful library developed by python foundation. Scipy is a widely used library that provides extra support for our model. Pandas is great for data analysis.

F. Best option.

I need more time to use these functions so I can find which one best fits our need and time-efficient to make the decision.