# Topic:

Good morning everyone. my name is Jiaqian. I am a first-year master student. Today I would like to present my final project which is to predict process performance using ML technique.

The prediction of OS process performance can bring several benefits, including:

Early detection of performance issues

Improved resource utilization

Better capacity planning

Reduced downtime

Improved user experience

Cost savings

And my goal for this project is to use ML technique to predict the performance and behavior of processes on operating system

To start the project, I have done some research and project design, aka determine what is the big picture of this project and draw a rough graph of how to accomplish my goal.

So here are my brief ideas of how to do it.

I plan to collect large amount of data related to process on both Linus and mac book. I need to gather data related to various performance metrics and system configurations. Once I have gathered the data, I need to preprocess data and make it ready to apply machine learning algorithms to build and train a predictive model that can predict processor performance. the data will be divided into trained data and test data. We can use regression algorithms such as linear regression, decision tree regression, or random forest regression to predict the performance. I can also use classification algorithms such as decision trees or neural networks to classify the performance as good or bad based on the input data. So sounds straightforward.

After that, I started to fill the details of my planning structure. Here is My progress guideline. I have divided the project into couple steps. There are Project design, Kernel module writing in c language, Collecting data on Linux and MacBook, Preprocess data, Build ml model for prediction, and last but not least create test cases to ensure the data we collect are accurate.

As you see, I have marked some of steps done and will provide demos for kernel module, data collecting and small part of implementaton of ML model in the next few slides. Besides, I have attached ed my current repository structure. It may change slightly since I haven’t 100% completed this project on coding part. Most of code are categorized into folders. For c language we need to compile the file to make it executable on Linux. So, I have created couple make file to make the work of typing command into terminal in an easier way to handle.

Steps:

Next I am going to talk about details for each steps

The first one is about kernel module. I have written a kernel module which help collects the data in Linux system. The goal for this kernel model is to log information related to processes. The kernel module itself will be installed in Linux.

Click on the demo video. Proclog is the kernel module file I built. As we see, we exec the make file for compiling the file and it will generate couple files with same name but different file types. The one with .ko is the one used to install in Linux system. When ko file is installed successfully, a msg will be shown to remind the user. ~~when we dump the kernel print by command `dmesg`~~. Meanwhile, ~~A kernel function called proc\_dir\_entry is created in folder /proc/ and it is~~ a virtual file called log\_file will be stored in /proc/ folder. In the demo I have dumped the first 10 processes which are logged into this entry. This log\_file is unwritable by user but only written by kernel module. As the linux environment keeps going on, the file is updated constantly.

Second step is to collect data from both mac book and Linux. Since it is dangerous for kernel to write an actual file so that an additional c file is needed to export content of virtual file into actual file. Because the log\_file keeps updated since the kernel module is installed, there are couple gigabytes of data in the virtual file as time goes on. We would like to make the export work easier by letting user input the number of lines of records to limit the export of data, or the reading will never reach to the end. ~~In the process we have done some data preprocessing by replacing the whitespace into comma which can be recognized by csv format.~~ it is similar step for collecting data on MacBook but I apply python on MacBook because it is easier to implement it on macOS. Python has its library called psutil to retrieve process data, which make it relative easier to collect data on MacBook. Because I didn’t install kernel module on MacBook to continuously collect data, I have set up an interval to let the user collect data by their choice. The program can retrieve process every interval so that the process may have some updated information to be stored into csv file. Besides, I have answered the question I have raised from my project proposal.

Third step,

I have used python to preprocess some data by dropping unnecessary data and filling the missing data in rows. I have chosen linear regression algorithm to build the ML model which has major uses of determining the strength of predictors, forecasting an effect, and trend forecasting. And I have chosen random forest regressor too. It is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. In the demo I have chosen each process kernel time as y and the other numeric columns as X. From both linear regression model and random forest regressor model, I have got over 96% accuracy between predicted value and actual values. I have applied Mean Squared Error (MSE) metric which is the most commonly used for evaluating a model. Also, I have answered the question I have raised from my project proposal.

# ~~What is Linear Regression?~~

~~Linear regression is a basic and commonly used type of predictive analysis.  The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?  (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?  These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.  The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.~~

~~Naming the Variables.  There are many names for a regression’s dependent variable.  It may be called an outcome variable, criterion variable, endogenous variable, or regressand.  The independent variables can be called exogenous variables, predictor variables, or regressors.~~

Graphical user interface, text, application

Description automatically generated

In addition, I have also encountered some challenges for this project.

Firstly, there are Limited resources and various restriction on building Kernel module. It is not easy to debug on kernel module. The error msg does not provide sufficient information for developer to locate the error. And as I mentioned earlier, an additional c file is needed to export content of virtual file into actual file. Thus it is Not efficient enough to collect linux data in this way. Compared with the data collecting approach I applied on macbook, it still has some advantage that the data is recorded constantly on virtual file. I can still choose to export large gigabytes of data without missing consistent data info since the kernel module is working.

Secondly, when I use python some data are not available such as io counter and net counter for each process. Disk io and network is can contribute to performance of processes too.

Besides, I have some plan for Future work.

1. In the next few days, I plan to Modify and wrap up the ML prediction
2. Create more test cases make sure prediction reliable
3. Find the Answer for my last question from project proposal: Does our model provide accurate prediction of performance? WIP to build test cases to verify.

That’s the end of my presentation. Any questrios or feedback?