AMOD 5610H Technical Feasibility

Name – Sahil Nagpal (0670042)

Partner Name – Pramina Patil (0663967)

In the technical feasibility stage students will need to demonstrate they have a working set of hardware and software, and can accomplish key challenges of the project

HardwareIn this section demonstrate that you can get the dataset loaded onto hardware, and that the hardware works (hopefully this part is trivial).   
If you have multiple group members, make sure you are demonstrating that each group member has access to what they need to work on the project – that might be everyone working from one server, or everyone having their own copy of the data. What is your plan if the hardware fails?

|  |
| --- |
| **Fill in this box:**   1. **Proof you have access to (some) data on real hardware.**   The dataset is downloaded from the anticipated website which has specifically allotted absenteeism data. The dataset is directly downloaded and loaded onto the local machine under a certain directory. The data as of now is present on the local hard disk and can easily be accessible to perform the processing.     1. **Show how your team connects to that data**   Since working in the group, the data sharing is an important part (especially when we are working simultaneously on same dataset). In order to share the data among the teammates we are using the cloud service (Amazon Web Service – Simple Storage Service (S3)). We have created the bucket under the amazon s3 and placed the dataset inside that bucket. From this bucket the teammates can download the data and can place the pre-process at the same location.  Amazon Simple Storage Service is storage for the Internet. It is planned to create web-scale computing easier for developers. Amazon S3 incorporates a basic web administrations interface merely can utilize to store and recover any sum of information, at any time, from anyplace on the net. It gives any designer access to the same exceedingly adaptable, solid, quick, reasonable information capacity framework that Amazon employments to run its own global network of web sites. The service looks towards to maximizing benefits of scale and to pass those benefits on to designers. |
| 1. **Discuss plan for what to do if hardware fails**   Chances are hardware can nosedive or disrupted at any time. To overcome the problem of the failure , our plan is to perform the processing on the cloud and store the resultant on the cloud (Amazon Web Services). The cloud service provides robustness of the hardware and provide the maximum reliability. Also , to maintain the hardware reliability we have taken the robust configuration machine on the cloud. For the data processing we have AWS Sagemaker cluster with 32 GB RAM and Intel Octa core processor with 1 TB SSD.  **What if the Hardware fails in between**  When the hardware fails or nosedive , the cloud service automatically boots up a new cluster (machine) which will handle the data and process failure and can restart the process again. For the storage , the AWS S3 maintains a copy of data on multiple availability zone which provides maximum availability. |

SoftwareAs with above, show that you have the most critical software you anticipate needing, that it’s installed and configured correctly, accessible to all relevant group members. If possible, you should also show that your software can, where applicable be connected to your actual dataset (it can load the data for example). E.g. An SQL Database holding data, and programming in Java, you should show that you can load data into the database and can query the database using Java/JDBC.

|  |
| --- |
| **Fill in this box  1. Software for accessing the data (this may be the same as in the hardware question)**  The software stack is always a food for thought. In order to make certain progress for this project we highly need a python as platform as a service. Also we need the Jupyter notebook as the software as a service which act as integrated development environment. For the data visualization purpose we highly need the interactive data visualization tool – Tableau. Also to maintain the code for the continuous changes and continuous deployment we need the Github software as the infrastructure as a service. Using the python and the inbuilt libraries and packages present in the python for instance : Pandas , Numpy and to name a few we can load our data and can perform the processing . These packages can help us to connect with our data.  **2. Explain your software stack at this point**  **Programming Language** – In order to build this project , we have chooses python (python 3.6) as primary programming language.  **Cloud Technology** – The cloud technology we are going to use is Amazon Web Services (Sagemaker , Simple storage service (S3)). The cloud computing as service helps us to built and deploy the project in more robust and precise way.  **Integrated Development Environment** – Jupyter Notebook is used as an Integrated development environment.  **Data Visualization** – For the purpose of the data visualization we are explicitly using the interactive data visualization tool – Tableau. Through this we are going to compare and define the different aspects and parameters which are built using the model.  **Version Control System** – In order to maintain the code version and to provide the continuous deployment we are using the GitHub open source.  **Libraries** – Pandas Numpy ,Scikitlearn,Seaborn,Matplotlib,Seaborn,Tensorflow and to name a few. |

# Sample of trivial processing/analysis of data

We’re not expecting the project to do anything here, but prove that you can actually do work on the data. E.g. you have a database, you can query it, you can get the results and then the results can be displayed. The point here is to show that you understand the problem, not that you’ve solved it.

|  |
| --- |
| **Fill in this box**  **Explain what the analysis of your data looks like – either show a sample of how a single row/document/whatever gets analysed, or show how the output is evaluated or essentially something that shows what the analysis actually is**  We first load the data using a Pandas    **Pre-processing being performed on data**  Since most of the data we are working contains the integer type of columns , since we need to perform some categorical operations , we have created the dummy variables for that. The dummy variables needs to be created for the ‘Reason of Absence’ Column. Since this depicts the actual reason of the absence.  Dummy Variables : an explanatory binary variable that equals 1 if the certain categorical effect is present and that equals 0 if that is same effect is absent.    The data we are dealing consist of the reason and to get the precise results we need to get the cluster of the reasons. The clustering here helps us to categorize the type of situation. So for that we have created the groups. The pre-processing is being done on the reason columns which is not present in the clustered way. |